Transformational Networks

Robert Frank and Donald Mathis

Department of Cognitive Science
Johns Hopkins University

It is a age-old observation that the sentences of human languages exhibit hierarchical organization, and that this organization is implicated in the mapping being sentence types. Thus, passives are related to actives through the displacement of noun phrase constituents (among other things), and interrogatives are related to declaratives through the fronting of an auxiliary verb that follows the noun phrase subject constituent. How and why do language learners derive structural generalizations about the patterns of their languages? Chomsky's (1975) Argument from the Poverty of the Stimulus (APS) starts from the premise that the relevant data to distinguish between structural and linear generalizations are absent from the learner's input. As a result, the only explanation for the structural basis of language must come from an innate learning bias, which Chomsky argues takes the form of a template for grammatical rules. Though the precise nature of the innate bias has evolved as linguistic theory has developed, the idea that there is a language-specific bias for hierarchical representations has remained constant.

The APS has recently come under fire. On the one hand, Pullum and Scholz (2002) have disputed the degree to which the stimulus is actually impoverished, finding examples of . Yet it remains an open question whether the infrequent presence of examples like (1), which distinguish a structure sensitive generalization for question formation (i.e., front the *main* auxiliary verb) from a linear sensitive generalization (i.e., front the *first* auxiliary verb), is sufficient to drive successful learning (Legate and Yang, 2002).

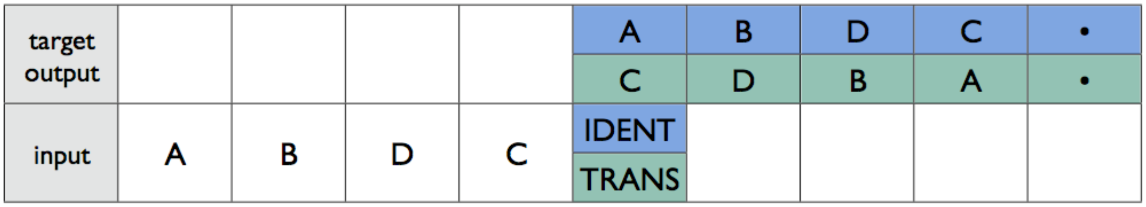
(1) Is the bird that is singing lonely?

Lewis and Elman (2001) stage a more direct assault on the APS, arguing that even in the absence of examples like (1), learners without any innate grammatical bias can nonetheless induce a structure sensitive generalization for question formation. Specifically, they trained a Simple Recurrent Network (SRN) to perform the task of word prediction on a variety of declarative and interrogative sentence types, withholding examples of the form in (1). In prior work, Elman (1991) had shown that SRNs exhibit sensitivity during the word prediction task to the non-local dependencies involved in subject-verb agreement, and on that basis he argued that they induce a hierarchical representation of the sentence. Lewis and Elman demonstrate that the network they trained generalizes in an apparently structure sensitive fashion when tested on cases like (1): at the relative pronoun *that*, the network predicts the occurrence of an auxiliary verb, and at the end of the relative clause it fails to predict an auxiliary.

There are a number of reasons for skepticism, however. First of all, there is little reason to believe that Lewis and Elman's network represents the relationship between the declarative and interrogative forms of a sentence (or alternatively between the fronted and canonical positions of the auxiliary verb) as such knowledge is unnecessary for the prediction task. Yet the question of structure sensitivity arises only in the context of this relation. As a result this simulation simply doesn't bear on whether an innate structure sensitive bias is necessary. Secondly, as Reali and Christiansen (2005) show, the distinction between the grammatical (1) and its non-structure sensitive counterpart (2) can be predicted using a bigram language model.

(2) Is the bird that singing is lonely?

However, as Kam et al. (2005) demonstrate, as soon as one expands the empirical domain slightly, bigram statistics are no longer sufficient to distinguish between the structure sensitive and linear sensitive patterns. It is therefore possible that Lewis and Elman's network is achieving its success through simple means, which will not generalize beyond their original experiment.

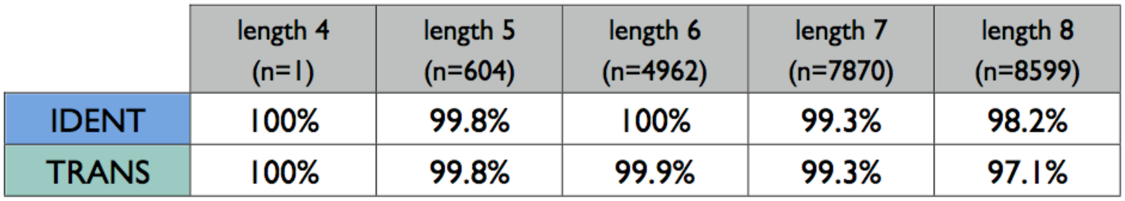


target output					A	B	D	C	•
					C	D	B	A	•
input	A	B	D	C	IDENT TRANS				

Figure 1: Training regimen for reversal network

To approach the question of structure dependence more directly, we moved away from the task of word prediction and focused instead the ability of a neural network to induce the kind of grammatical mappings that were the basis of Chomsky's original argument, namely structure-dependent transformations. There have been a number of previous attempts to get networks to learn structure sensitive mappings (Chalmers, 1990; Niklasson and van Gelder, 1994; Neumann 2002). However all of these works share the assumption that the network is presented with a representation that in some manner encodes the hierarchical structure of the input. Given such an input, it is the task of the network to learn a mapping between this representation of hierarchical structure and another. Yet the situation of language learning does not present a learner with hierarchical syntactic structure. Leaving aside the possible role of prosodic information, any hierarchical structure that is necessary to account for syntactic regularities must be imposed by the learner. Since SRNs have been touted as an instance of a system that can induce hierarchical structure from sequential input, we aimed to investigate their effectiveness in learning to transform sentences from one grammatical form to another.

Although past studies of SRNs have made great use of their ability to accept temporally ordered input, allowing them to take unboundedly long sentences as input, that work has not addressed the question of unbounded outputs of the sort that must be allowed as possible outputs of a grammatical transformation. Botvinick and Plaut (2006) provide a simple and elegant way to do this in their studies of short-term memory for serial order. Botvinick and Plaut demonstrated that when given a sequence of letters as input, an SRN can be trained, upon the presentation of a recall cue, to output the input sequence one element at a time. In order to assess the limits of this ability, we presented an SRN with a somewhat more complex task: instead of a single recall cue that triggered the identical sequence as output, we introduced an additional cue whose target output was a transformation of the original sequence. In the simulation, the input sequences (of which there were 72,000) were drawn from a set of four symbols {a, b, c, d}, and varied in length from 1 to 8, in equal numbers. Training consisted of the presentation of one of these sequences one symbol at a time, with no target output, followed the presentation of one of two recall cues (IDENT or TRANS) for a single time step, which triggered the target output sequence that was either the identity or reversal of the original. This is depicted in Figure 1. The input and output layers of the network contained 6 units, and the output contained 5 units, and these were used for localist representations of the input and output symbols. The hidden and context layers contained 100 hidden units. All units but the outputs used sigmoid activation functions, while the outputs used a soft-max activation function, so that activation was interpretable as the network's assessment of the probability of a particular unit as output. This network was trained for 120,000 weight updates with the Backpropogation Through Time algorithm, using a cross-entropy error function, a batch size of 50 examples, and initial random weights in the range [-.1,+.1]. As seen in Figure 2, this network is extraordinarily successful when tested on novel sequences. An output sequence was judged as correct only if each of the targets was the most active output unit at the appropriate time step.



	length 4 (n=1)	length 5 (n=604)	length 6 (n=4962)	length 7 (n=7870)	length 8 (n=8599)
IDENT	100%	99.8%	100%	99.3%	98.2%
TRANS	100%	99.8%	99.9%	99.3%	97.1%

Figure 2: Accuracy of reversal network on novel sequences

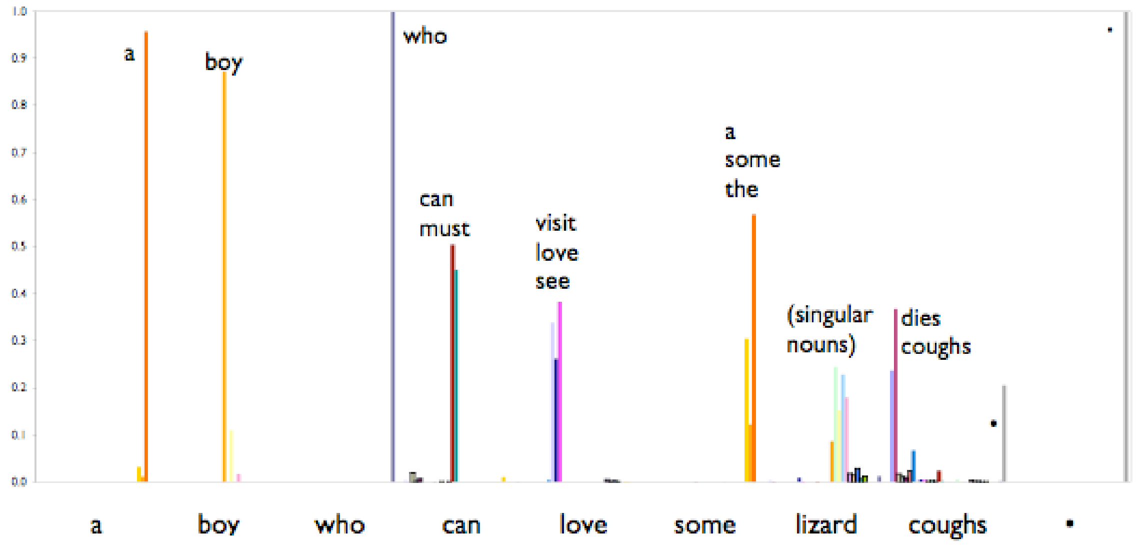


Figure 3: Network output for declaratives

Impressive as it is, the network's success at performing the reversal task does not guarantee success in a grammatical transformation task. As a transformation, reversal does not require sensitivity to any sort of structure in the sequence. We therefore attempted to train an SRN to perform a grammatical transformation on an input sentence, specifically the mapping from declarative to interrogative sentences. Following the design in Lewis and Elman (2001), we trained the network using simple sentences with both transitive and intransitive verbs, with and without auxiliary verbs, subject-verb agreement, and recursive modification of noun phrases by prepositional phrases and relative clauses. All inputs to the network were declarative sentences, while the target output sequence either consisted of the identical declarative (when triggered by a DECL cue) or an analogous interrogative (when triggered by a QUEST cue). Because the vocabulary of the network was larger, the number of input and output units was increased to 34 to allow for localist representations of the entire vocabulary. The number of hidden and context units remained at 100. The training data consisted of 100,000 stochastically generated sentence inputs (average length of 5.54 words, ≈15% including a prepositional phrase modifier, ≈8% with a relative clause modifier), with half of these were followed by a DECL cue, and the other half followed by a QUEST cue, with the appropriate declarative or interrogative sentence as the target output sequence after this point. In order to replicate Lewis and Elman's scenario, we withheld from training one class of training examples: those with a relative-clause-modified subject and a QUEST recall cue. This meant that although the network was exposed to sequences in the input with subjects modified by relative clauses, it was instructed on how to form questions from them. If the network had represented its knowledge of the question transformation in a structure sensitive fashion, so that it encoded a generalization about all kinds of noun phases in subject position, we should expect to find generalization to this held-out example type. In contrast, if the network has represented the question transformation as a mapping between linear sequences of

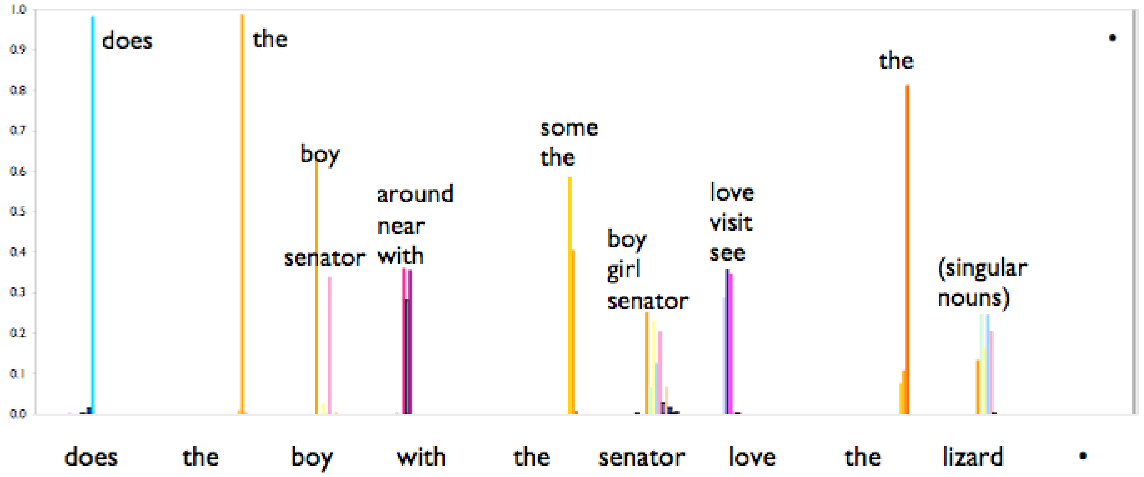


Figure 4: Network output for interrogatives

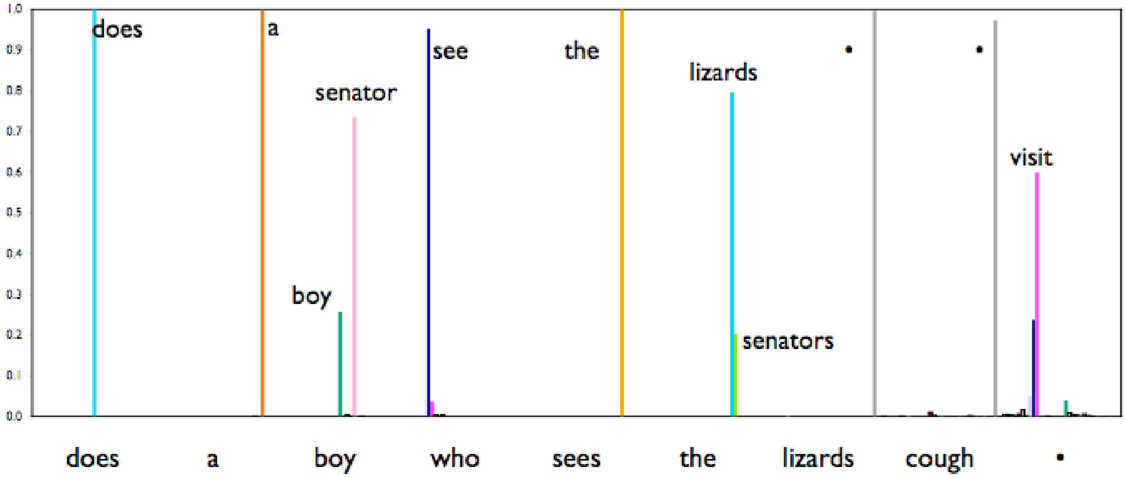


Figure 5: Network output for withheld interrogative

words, the absence of such a pairing in the training data would prevent the network from generalizing to the novel example type.

For the sentence types on which the network was trained, its performance was highly accurate, with the only errors being the substitution of one word by another within the same grammatical class. Examples of the network's outputs as compared to the target outputs for a declarative and interrogative output sentence is shown in Figures 3 and 4 (along the horizontal axis are the target words, vertical bars represent activation of lexical outputs). In contrast, the network was unsuccessful for its interrogative outputs for sentences of the type on which training was withheld. Indeed, the sequence of words output by the network never matched the target, even abstracting away from errors with word class. An example output is shown in Figure 5. Instead, the network's outputs almost always corresponded to output sequences of a type on which the network was trained. These erroneous outputs were not however random. Typically, though not always, they were well-formed questions of some sort, and they generally preserved one of two properties of the input sequence (or both):subsequences of lexical items from the input, or sentence length (the example in Figure 5 preserves the former property). We re-ran this simulation under a variety of conditions, changing the number of hidden units, batch size, learning rate, and vocabulary size, with the same qualitative result emerging in all cases.

The inability of the network to produce a question of the appropriate structure suggests that the network has not induced an abstract structural generalization about question formation that cuts across different instances of noun phrases in subject position. However, the fact that the network does not produce an output that is interpretable as the output of any coherent structural or linear transformation makes it difficult to determine just what sort of knowledge the network has acquired. Indeed, we suspect that the network induces some sort of "output grammar" on the basis of the sequences that it has been trained to output, and this grammar constrains the possible network outputs, even if the network's internal representation could be taken, in some sense, to represent the correct output of the transformation. We are investigating this possibility in ongoing work by considering whether other training regimens might allow the network to produce such output forms.

In spite of the possible presence of an output grammar, there is one way in which we might be able to diagnose the structure sensitivity of the network's knowledge. Consider, for instance, a sentence like the following:

(3) A boy who can love some lizards must cough.

We already know that our network will fail to produce an interrogative corresponding to this sentence. At the very first word of the output, however, the network will need to produce an auxiliary verb of some sort. If the network's transformation of (3) into a question is structurally-based, we should expect to find that the first word in its output will be *must*. In contrast, if the generalization is linearly-based, the first output will be *can*. A third possibility is that the network has learned a default

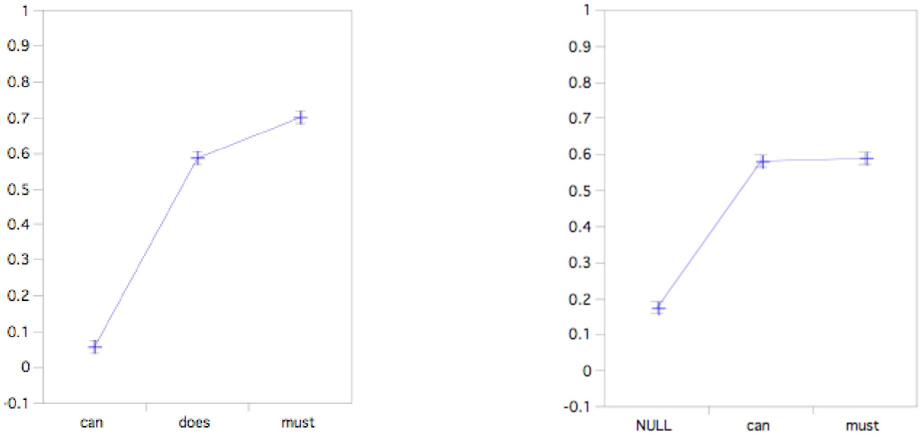


Figure 6: Activation of target auxiliary. Left graph gives activation as the target varies, Right graph gives activation on as the linearly first auxiliary (within the relative clause) varies.

of sorts, so that its output does not depend on the lexical content of the auxiliaries in the input. We can therefore find out something about the structure dependence of the network's behavior by observing the activation of this first word.[1] To test which of these possibilities characterized the network's behavior, we tested the network on a set of 161 sentences, all containing relative-clause-modified subjects, in which the bearer of verbal inflection in the main and relative clauses systematically varied among two modal verbs (*can* and *must*) and the main verb. We found that the mean activation of the correct auxiliary verb at the point immediately following the recall cue was .45. Average activation however varied by the target auxiliary: as illustrated in the left graph in Figure 6, the network's average activation when the target was *can* was virtually 0, while it was much higher for the other two possibilities. The right graph shows that the network's success in producing the correct auxiliary also varied depending upon the identity of the auxiliary in the relative clause modifying the subject (the linearly first auxiliary): when there was no auxiliary within the relative clause, indicated by the label NULL in the graph, as in a sentence like *a boy who loves some lizards can cough*, the network was quite unlikely to produce the correct auxiliary to start the question, but when the relative clause contained one of the modal verbs, the network was much more likely to correctly produce the correct auxiliary. Subsequent replications of this simulation, with different initial random weights, yielded qualitatively similar results.

The network's success in correctly producing the auxiliary verbs *does* and *must* does point to some sort of structural dependence in its knowledge of question formation. However, the network seems unable to put aside irrelevant non-structural factors, such as the identity of the auxiliary in the relative clause, in the formulation of its generalization concerning question formation. Concerning the first of these, it is possible that it is an accurate reflection of the path of child language acquisition. Santelman et al. (2002) *inter alia* have found that children vary in their success in producing correctly inverted questions depending upon the identity of the auxiliary verb, though they found worse performance with *do* than with modals. We leave for future work the question of whether this pattern might arise in a training set with more realistic distributions among the types of auxiliaries. Concerning the sensitivity to the linearly first auxiliary, we are unaware of evidence that children or adults show a similar pattern. We note, however, that we have found a similar inability of SRNs to attend to linearly-based generalizations in on-going work on the induction of anaphora. Contrary to what is sometimes assumed, then, the difficulty SRNs have in inducing grammatical generalizations does not reside in identifying structurally-based generalizations, but rather in ignoring linearly-based ones. Since it was precisely the ability to put aside such non-structural generalizations that was at the crux of Chomsky's APS, we contend that the argument still stands.

Acknowledgments

This work has been supported by NSF grant SBR-0446929. For help in running the simulations and useful discussion, we are grateful to Ebony Gussine, John Stowe, Manny Vindiola, and especially Esteban Buz. We have also benefited from the comments of colleagues at UCLA, University of

[1] Thanks to Bill Idsardi for this suggestion.

Delaware, AMLAP, and University of Maryland, where we have been fortunate to have the opportunity to present some of this work.

References

Botvinick, Matthew and Plaut, David C. 2006. Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, 113:201–233.

Chalmers, David J. 1990. Syntactic transformations on distributed representations. *Connection Science*, 2(1& 2):53–62.

Chomsky, Noam. 1975. *Reflections on Language*. New York: Pantheon.

Elman, Jeffrey L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

Kam, Xuan.-Nga Cao, Stoyeshka, Iglika, Tornyova, Lidiya, Sakas, William G., and Fodor, Janet D. 2005. Statistics vs. UG in language acquisition: Does a bigram analysis predict auxiliary inversion? In *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition*, pages 69–71, Ann Arbor. Association for Computational Linguistics.

Legate, Julie A. and Yang, Charles D. 2002. Empirical re-assessment of stimulus poverty arguments. *Linguistic Review*, 19:151–162.

Lewis, John D., and Elman, Jeffrey L. 2001. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th Annual Boston University Conference on Language Development*.

Neumann, Jane. 2002. Learning the systematic transformation of holographic reduced representations. *Cognitive Systems Research*, 3(2):227–235.

Niklasson, Lars F. and van Gelder, Tim. 1994. On being systematically connectionist. *Mind and Language*, 9:288–302.

Pullum, Geoffrey K. and Scholz, Barbara C. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19:9–50.

Reali, Florencia, and Christiansen, Morten H.. 2005. Uncovering the richness of the stimulus: Structure dependence and statistical evidence. *Cognitive Science*, 29:1007–1028.

Santelmann, Lynn, Berk, Samantha, Austin, Jennifer, Somashek, Shamitha, and Lust, Barbara. 2002. Continuity and development in the acquisition of inversion in yes/no questions: dissociating movement and inflection. *Journal of Child Language*, 29:813–842.