

Question Formation, Neural Networks and the Poverty of the Stimulus

Robert Frank, Donald Mathis, Ebonye Gussine, John Stowe, and Manny Vindiola

DEPARTMENT OF COGNITIVE SCIENCE
JOHNS HOPKINS UNIVERSITY

1 Structure Dependence

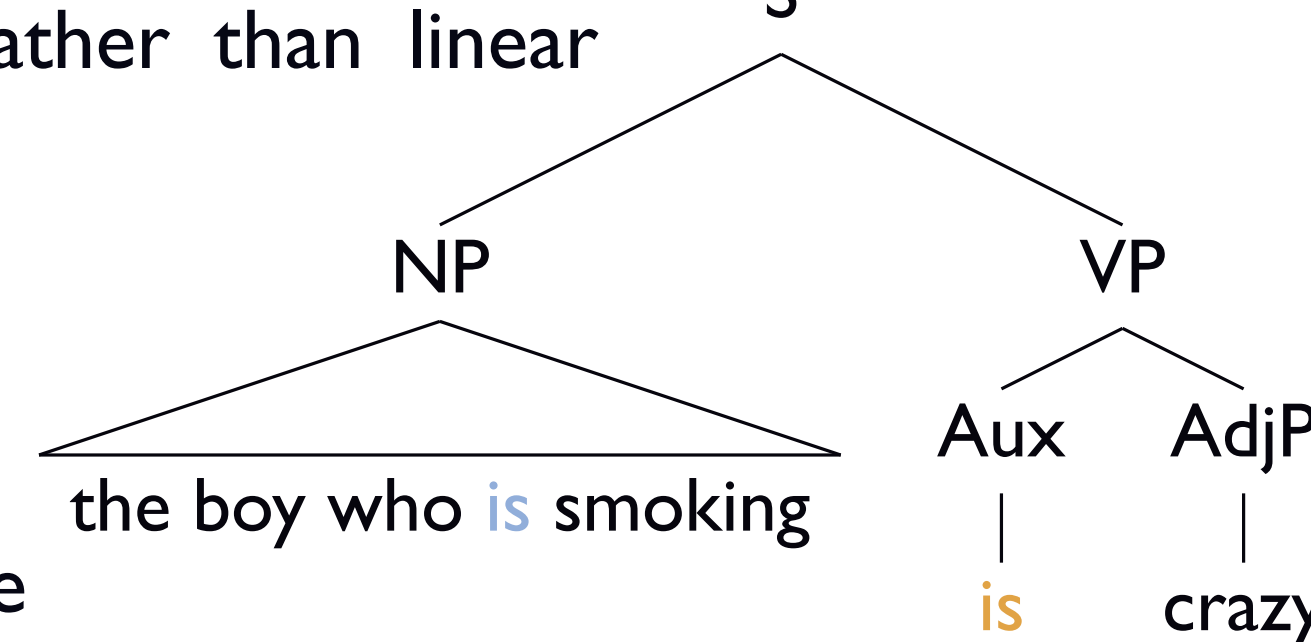
Chomsky's (1975) poverty of the stimulus (POS) argument:

1. The rule for forming questions in English is sensitive to structure ("move the main auxiliary") rather than linear order ("move the first auxiliary").

- (1) *Is* the boy who *is* smoking crazy?
- (2) **Is* the boy who smoking *is* crazy?

2. The relevant evidence that distinguishes these two possibilities is absent from the child's input (but cf. Pullum & Scholz 2002).

3. There must be an innate bias in favor of structure sensitive grammatical generalizations.

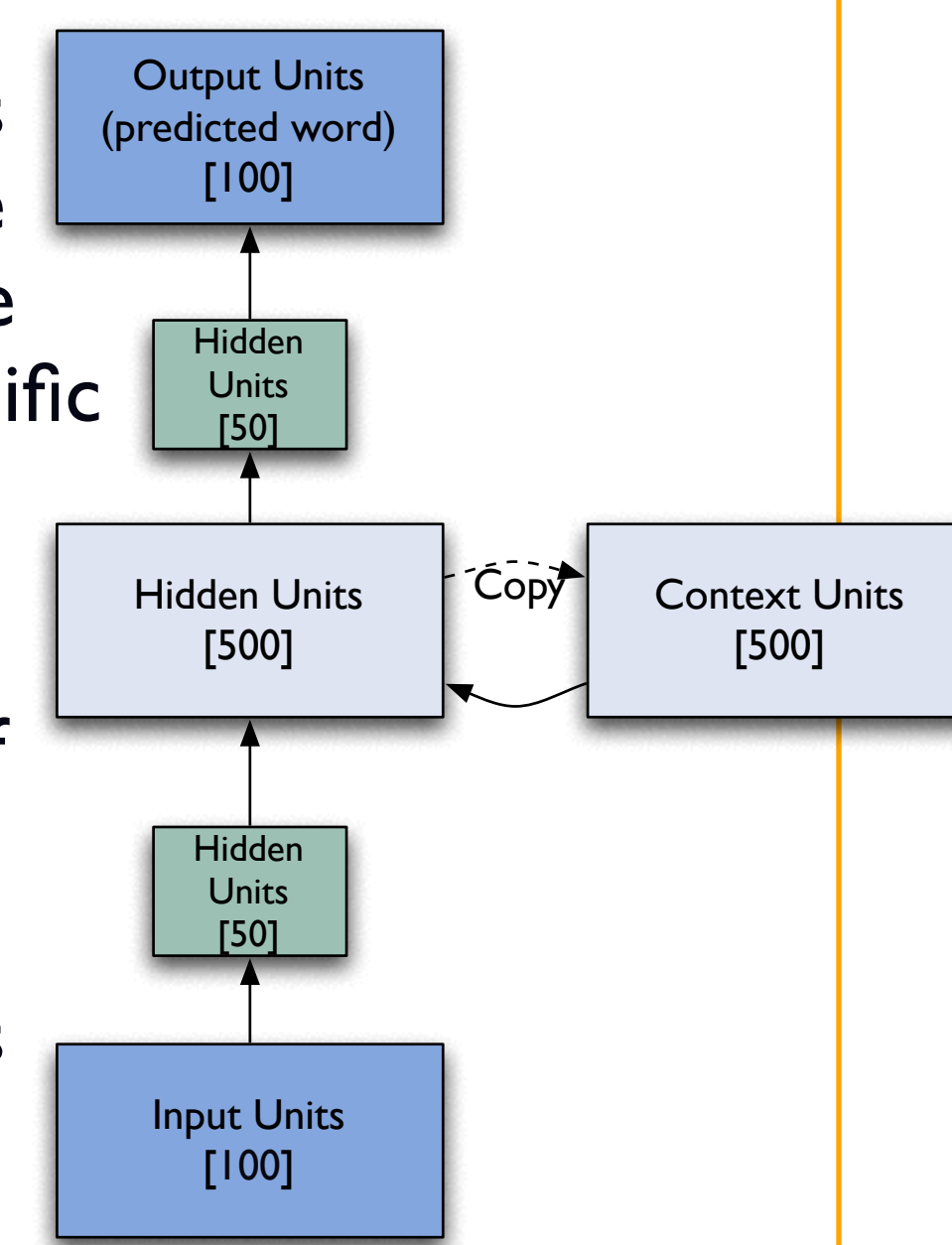


2 The Connectionist Reply

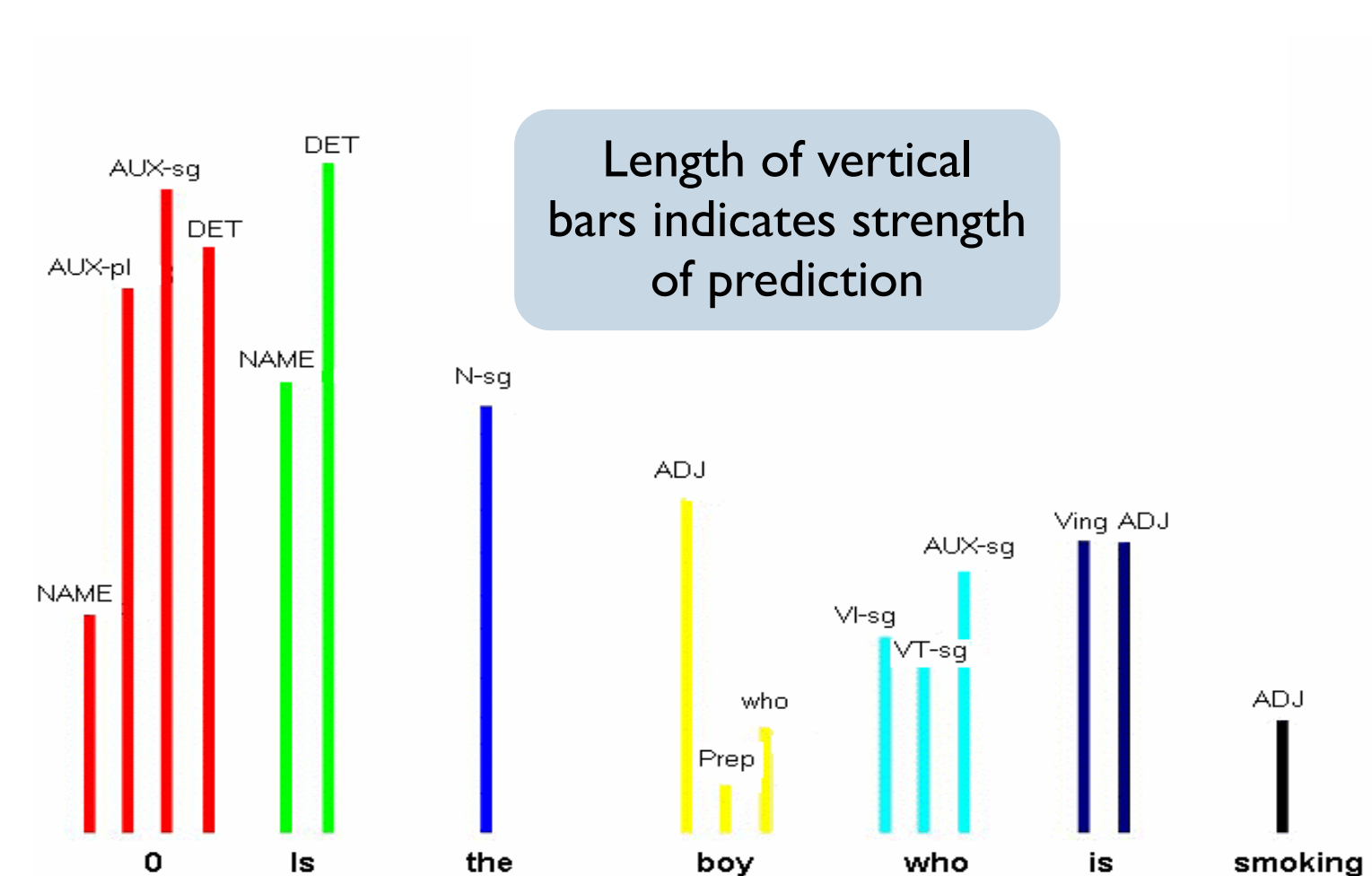
Lewis and Elman (2001): Chomsky's POS argument ignores the rich statistical information present in the input. Simple recurrent networks (SRNs) can induce structure sensitive generalizations without any innate language-specific machinery.

The simulation:

- Train an SRN to do word prediction over a corpus of English declarative and interrogative sentences.
- Words are locally represented at input and output.
- During training withhold interrogative sentences whose subjects are modified by relative clauses.
- Test trained network on withheld sentence type.



The results:



At relative pronoun *who*, network predicts an auxiliary, suggesting that it has not induced a linear rule for auxiliary fronting. At end of relative clause (at *smoking*), network does not predict an auxiliary (in this sentence), indicating that it has generalized to this case.

What has this network really learned?

We are grateful to the National Science Foundation for their financial support of this work in the form of grant SBR-0446929.

Puzzle 1: the bigram objection

Real and Christiansen (2005) show that examples (1) and (2) can be reliably distinguished through a simple bigram probability model.

$$p(\text{is the boy who is smoking crazy}) = p(\text{is})p(\text{the|is})p(\text{boy|the})p(\text{who|boy})p(\text{is|who})p(\text{smoking|is})p(\text{crazy|smoking})$$

$$p(\text{is the boy who smoking is crazy}) = p(\text{is})p(\text{the|is})p(\text{boy|the})p(\text{who|boy})p(\text{smoking|who})p(\text{is|smoking})p(\text{crazy|is})$$

Kam et al. (2005) observe that it is one bigram, the high frequency "who is," which yields a significant probability difference between (1) and (2). In a more realistic subset of English without such distinctive bigrams, such a distinction is no longer possible (*Is the wagon your sister is pushing red?* vs. *Is the wagon your sister pushing is red?*). Is the SRN exploiting such a linear bigram model, rather than hierarchical structure, to achieve its success?

Puzzle 2: the correspondence objection

The prediction task does not demonstrate knowledge of the mapping between declarative and interrogative sentences. It is only in the context of this correspondence that the question of structure dependence arises. Otherwise, the two sentence types may be treated independently.

3 Transformational Networks

To represent the relationship between sentence types, we trained an SRN to perform a generation task. Following Botvinick and Plaut (2006), we trained an SRN to output a sequence after an output signal: We used two such signals: IDENT required an identical sequence to be output; TRANS required a transformed sequence (e.g., reversal)

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

Simulation details:

- SRN with 100 hidden and context units, 7 input and output units
- Training data included 72K examples from {a,b,c,d}⁺ of length 1 through 8 (in equal numbers) with no target output, followed by one time step of IDENT or TRANS, followed by target output which was the original or reversed sequence.
- Training with BPTT, cross-entropy error function, batch size 50, initial weights in [-.01,+.01], 120K weight updates

Accuracy on held-out strings:

	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%

This technique works for simple formal transformations. What about linguistically relevant transformations, like question formation?

4 Simulating Question Formation

target					the	dog	will	trip	•
output					will	the	dog	trip	•
input	the	dog	will	trip	DECL				
					QUEST				

Linguistic Domain:

- Subcategorization: transitive and intransitive verbs
- Modal verbs and *do-support*
- Subject-verb and determiner-noun agreement
- Recursive modification: PP and relative clause modifiers

	Simple subjects	PP-modified subjects	RC-modified subjects
Declarative	Some boys should visit the chinchilla who can laugh.	The girl near the senator can cough.	A boy who can love some lizard coughs.
Interrogative	Will the dog trip?	Does the boy with the senator love the lizard?	Does the boy who sees the lizards cough?

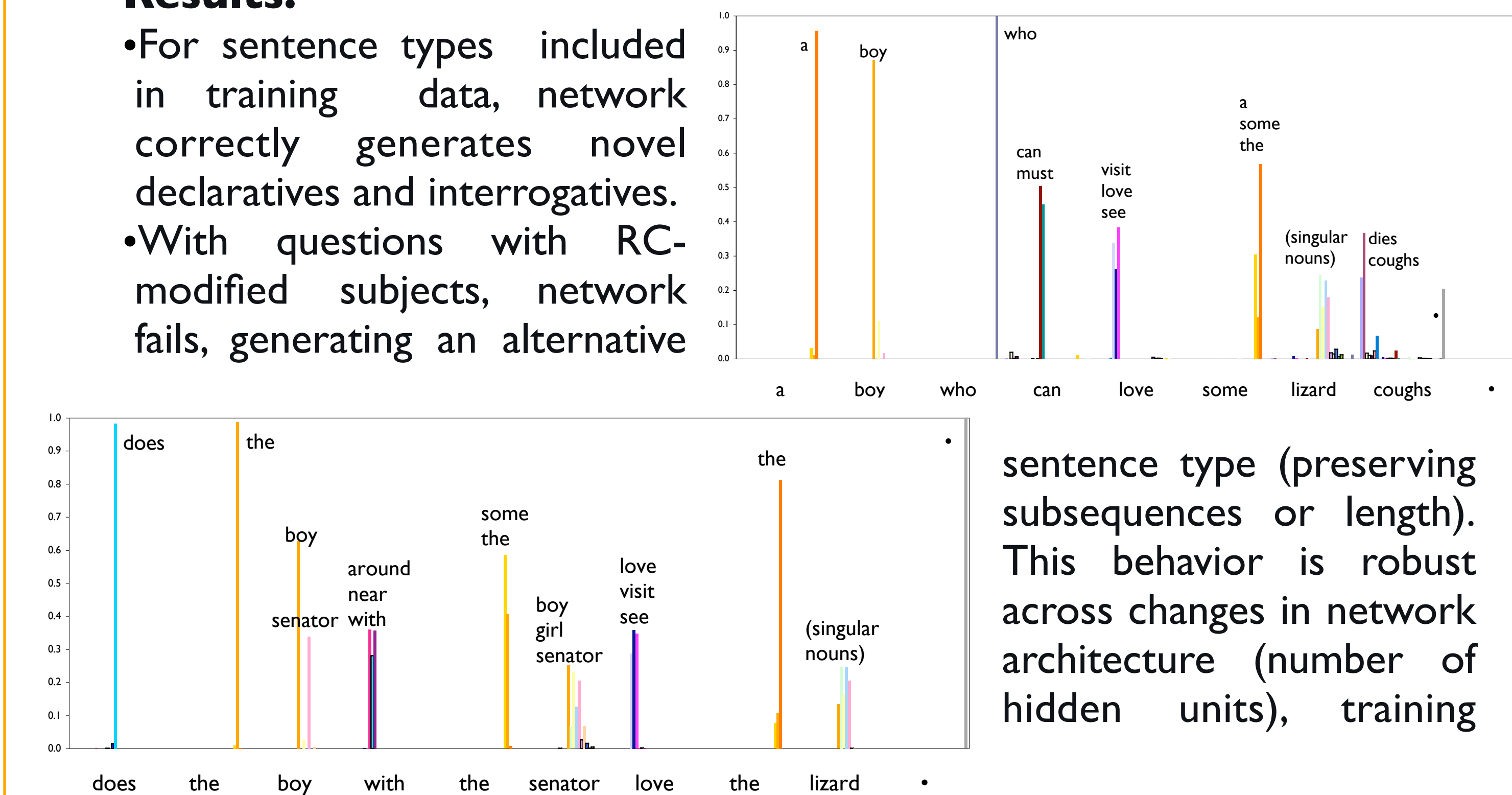
Held out from training

Simulation details:

- SRN with 100 hidden and context units, 34 input and output units
- Training data: 100K sentences generated stochastically (avg. length 5.54, ≈15% with PP, ≈8% with RC), half followed by D cue, half followed by Q cue, with appropriate declarative/interrogative target sequence.
- Training with BPTT, cross-entropy error function, batch size 50, initial weights in [-.01,+.01], 295K weight updates

Results:

- For sentence types included in training data, network correctly generates novel declaratives and interrogatives.
- With questions with RC-modified subjects, network fails, generating an alternative



sentence type (preserving subsequences or length). This behavior is robust across changes in network architecture (number of hidden units), training

regimen (batch vs. on-line training, learning rate, vocabulary size).

These results indicate that SRNs do not acquire a single abstract structural generalization governing question formation and therefore they do not at present undercut Chomsky's POS argument.

References

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

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

Kam, Xuân-Nga Cao, Iglia Stoyeshka, Lidiya Tornyova, William Gregory Sakas, and Janet D. Fodor. 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*, 69–71. Ann Arbor: Association for Computational Linguistics.

Lewis, John D., and Jeffrey L. Elman. 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*.

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

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