Language Acquisition

ISSN: 1048-9223 (Print) 1532-7817 (Online) Journal homepage: https://www.tandfonline.com/loi/hlac20

The Acquisition of Anaphora by Simple Recurrent Networks

Robert Frank , Donald Mathis & William Badecker

To cite this article: Robert Frank , Donald Mathis & William Badecker (2013) The Acquisition of Anaphora by Simple Recurrent Networks, Language Acquisition, 20:3, 181-227, DOI: 10.1080/10489223.2013.796950

To link to this article: https://doi.org/10.1080/10489223.2013.796950

Language Acquisition, 20: 181–227, 2013

ISSN: 1048-9223 print / 1532-7817 online
DOI: 10.1080/10489223.2013.796950

Routledge
Taylor & Francis Group

ARTICLE

The Acquisition of Anaphora by Simple Recurrent Networks

Robert Frank
Yale University

Donald Mathis
Johns Hopkins University

William Badecker
University of Arizona

This article applies Simple Recurrent Networks (SRNs; Elman 1991, 1993) to the task of assigning an interpretation to reflexive and pronominal anaphora. This task demands more refined sensitivity to syntactic structure than has been previously explored. Measured quantitatively, SRNs perform quite well. However, the way in which they achieve such performance diverges in key respects from the target grammar: (i) linear N-V-reflexive/pronoun sequences affect the SRN's interpretations, even without a relevant structural relation, yielding errors unlike those made by humans; (ii) the SRN's representations distinguish sentence types, inhibiting structural generalization; (iii) the SRN's knowledge of the conditions on anaphoric dependencies fails to generalize to novel lexical items. These results have important consequences not only for the viability of SRNs as models of language learning but also for the systematicity of generalization in neural networks (Hadley 1994; Marcus 1998).

1. INTRODUCTION

Questions concerning the nature of human linguistic knowledge, the manner in which language is acquired, and the way in which language is processed have played a central role in shaping research in cognitive science. Though there have been and continue to be many disagreements, there are two fundamental results that are widely agreed upon. First, the wide variety of patterns that exist in the languages of the world are best characterized in terms of hierarchically organized structural descriptions (as evidenced, for example, by the great many syntactic processes

Correspondence should be sent to Robert Frank, Department of Linguistics, Yale University, 370 Temple Street, P.O. Box 208366, New Haven, CT 06520-8366. E-mail: bob.frank@yale.edu

that cause word order to depart from the canonical sequence of phrasal groupings in a language through the displacement of syntactic constituents). Secondly, although languages differ from one another in many ways, there appear to be certain dimensions along which they do not vary (e.g., in their adherence to structurally based restrictions on coreference, and in the kinds of locality constraints that govern phrasal displacement).

If such structural descriptions and grammatical invariants form the core of our linguistic knowledge, an important question arises as to how speakers come to possess such knowledge. Within the field of generative linguistics, the abstractness of this knowledge and its conceptual distance from the data available to the language learner has led researchers to the conclusion that humans possess an innately specified and finely structured language faculty, so-called Universal Grammar (UG). On this view, UG provides the inductive bias necessary to lead language learners to attend to structurally defined regularities and to draw similar conclusions even in the face of disparate and incomplete data, leading to the observed invariants. For many researchers, some generative linguists included, this picture remains unsatisfying. Without some sense of why or how a particular grammatical property might be innate, the theory is infused with the unappealing taste of stipulation, no matter how specific or universal the property can be shown to be.

Over the past 30 years, the connectionist paradigm has emerged as a promising alternative to the symbolic paradigm in a wide range of areas of cognitive theorizing. Connectionist approaches to the mind eschew abstract symbolic representations and rules in favor of subsymbolic computation by neuron-like units. Where symbolic approaches to cognition, and to language especially, have emphasized the role of nature—of innate structure—in the emergence of law-like regularities, connectionist research emphasizes nurture, using techniques for statistical induction on the data provided to a learner as a means of extracting generalizations. As Elman et al. (1996) have emphasized, the connectionist approach does not deny the importance of innately provided inductive bias; instead it locates such bias in the properties of neural architecture, for instance the number, type, and connectivity of neural units, rather than in innate domain-specific principles.[1]

This line of inquiry has obvious appeal, and part of this appeal is that the kinds of innate structure necessary under the connectionist perspective are seen by some as more biologically plausible than that which is required under the UG approach. However, plausibility at this level of description is not sufficient on its own. In order for a connectionist account of language learning to be convincing, it must be shown precisely how the sorts of structural regularities seen in natural languages arise from connectionist networks when they are endowed with an appropriate architecture and trained on realistic linguistic data. There has been some work in this connection in the domains of phonology and morphology, most famously the work spawned by Rumelhart and McClelland's (1986) model of the English past tense (see, for example, Pinker & Prince 1988, Pinker & Ullman 2002, and McClelland & Patterson 2002, for discussion). Yet, there has been a barrier to applying these methods to the problem of learning the structure of sentences: Any network must have a fixed bounded number of input or output units, while sentences can grow without bound. How then can sentences of unbounded length be represented as the input or output of a network?

Elman (1991) proposes a way of resolving this impasse using Simple Recurrent Networks (SRNs). Unlike traditional feed-forward networks in which the output of the network depends

[1] Other approaches to language learning marry statistical induction to a language-specific (UG) conception of inductive bias. See Yang (2004) and Pearl (2011) for representative examples. Such approaches, which we think are promising, hold out the possibility of reducing, though not eliminating, the explanatory burden placed on language-specific innate knowledge.

entirely upon the levels of activation at the input units (along with the network connectivity), SRNs include recurrent connections that record information about the previous state of the network. Consequently, a sentence can be fed to an SRN one word after another, and the state of the network can encode information about the preceding context. Elman proposes testing the ability of SRNs to induce a representation of sentence structure by training them to perform word prediction: The words in a sentence (or rather corpus of sentences) are fed to an SRN sequentially, and the SRN is trained to predict the next word in the sequence. Since the ability to make such predictions rests on an awareness of the sentence structure, success at this task provides one sort of evidence that the network has in fact induced information about such structure. Rodríguez, Wiles & Elman (1999) and Rodríguez (2001) have shown that an SRN trained in this fashion is capable of inducing the kind of structurally based generalizations representable by the class of context-free grammars.

Especially relevant to our current concerns is the work of Elman (1993), who explores the ability of SRNs to learn to do word prediction in the context of an interestingly complex fragment of English. Specifically, Elman focuses his attention on a corpus of simple English sentences that have two salient grammatical properties: number agreement between the subject and verb, and the possibility of relative clause modifiers attached to subject and object noun phrases. In the context of sentences like *the apple/apples was/were red*, subject-verb agreement can be resolved on the basis of linear relationships between adjacent words. However, in the presence of (unboundedly many) relative clauses between the subject and verb in examples like *the apple/apples that fell on the scientist who proposed gravity was/were red*, the subject-verb relationship can no longer be characterized in terms of linear relationships, because there is no fixed distance or finite set of distances between the words that must enter into an agreement relation. Instead, inducing the correct generalization about the subject-verb relation in this class of English sentences requires reference to hierarchically structured sentence representation, under which this relation is structurally uniform. Under certain training conditions Elman shows that an SRN can succeed at the prediction task for this class of sentences. On the basis of this result, Elman argues that learners need not come to the task of language learning with the language-specific predisposition to learn hierarchical grammatical structure. Instead, such structure emerges on the basis of the network's training from the input data itself.

This paradigm has been applied to other grammatical constructions, with similarly successful results. Lewis & Elman (2001), for instance, show that SRNs trained to do word prediction learn the conditions on subject-verb inversion, apparently inducing the correct structure-based rather than linear-order-based generalization (cf. Crain & Nakayama 1987). Rohde (1999a) demonstrates that the SRNs can acquire knowledge of the conditions under which *want to* can contract to *wanna*, conditions that have been argued to require reference to abstract entities such as traces of *wh*-movement.

These demonstrations are impressive and suggestive of the power of SRNs. Yet, they leave open a number of important questions concerning the adequacy of SRNs as models of language learning. First of all, the phenomena that have been studied to this point only begin to skim the surface of the rich range of structurally rooted generalizations that are seen in the grammars of individual languages. Secondly, there has been relatively little study of the precise manner in which the trained network analyzes its linguistic input and whether the acquired knowledge reflects the same sort of generalizations that human language learners have been shown to acquire.

This article lays out our first explorations of both of these issues, by focusing on the ability of SRNs to extract generalizations concerning the interpretation of anaphoric elements, specifically

pronouns and reflexives.[2] As we will outline in the next section, such interpretation is sensitive to fine details of syntactic structure that are not local in the same way as those underlying subject-verb agreement. Further, the richness of this domain allows us to probe the nature of the generalizations that the networks acquire and compare them to those acquired by human language learners.

Our choice of anaphoric interpretation as the domain of investigation is also motivated by the fact that the generalizations that need to be learned are both *lexically* and *structurally abstract*. By *lexically abstract* we mean that the conditions on the interpretation of a reflexive like *herself* are uniform across occurrences of the reflexive with any possible antecedent of the reflexive and do not vary from one individual, say Mary, to another, say Sue. As observed by Marcus (1998), such lexical abstractness of anaphoric interpretation can be taken to implicate the use of a rule containing a variable, one which may be instantiated by any structurally accessible antecedent.[3]

By *structurally abstract*, we mean that the language learners acquire generalizations about the possible interpretations for a reflexive element that cut across a variety of sentence contexts. For instance, the possibility of interpreting a reflexive direct object of a verb as coreferent with the subject does not in English or any other language we are aware of depend on whether the subject does or does not have a relative clause attached to it, whether the verb is in a simple or compound tense, or whether the sentence is negative or positive. Structural abstraction points again to the use of a variable-based rule, where the variable in question must match all of the possible structural contexts in which anaphoric interpretation is possible. In our study of the performance of SRNs in the task of anaphoric interpretation, then, we will then address the following pair of questions:

1. How lexically abstract is the knowledge SRNs learn? Are grammatical dependencies learned as general relations between constituents (or word classes) or as relations between specific words?
2. How structurally abstract are the generalizations that SRNs learn about grammatical dependencies? To what degree can SRNs abstract over different structures to form a unified generalization about a grammatical dependency?

Not only will the answers to these questions enable us to understand better how SRNs solve linguistic tasks, but they will also allow us to explore the ability of these networks to acquire systematic generalizations. Fodor & Pylyshyn (1988) first raised the question of whether neural networks could display truly systematic, combinatorial behavior, where the units of language or thought are freely combinable (e.g., the ability to process *The politician loves Mary* entails the ability to process *Mary loves the politician*). Hadley (1994, 2003) sharpens Fodor and Pylyshyn's systematicity question to be one about the kind of generalizations an agent can extract from a

[2]Joanisse & Seidenberg (2003) have also applied SRNs to the problem of anaphoric interpretation. Their focus, however, was quite different than ours: on modeling the effects of impaired working memory of phonology on performance in this domain. Perhaps because of this, they did not study in great detail the successes and failures of their intact network on the anaphora task.

[3]Note that lexical abstractness is not universally assumed for the acquisition of all grammatical processes. For instance, Tomasello (1992, 2000), among others, argues that subcategorization frames and argument structure alternations are learned on a verb-by-verb basis. See Fisher (2002) for a different view. Whatever the resolution of this debate, we take it to be uncontroversial that the acquisition of anaphora does not work in this way, and indeed it is conspicuous that to our knowledge such a lexically specific proposal has never been made in this domain.

limited set of data, defining a number of levels of systematicity that an agent may exhibit. Among these is the notion of *strong systematicity*, which is defined as follows:[4]

(1) An agent displays *strong systematicity* if it can process test sentences that use a significant fraction of its vocabulary in novel syntactic positions.

This notion corresponds roughly to the idea of lexical abstraction: Only if a network represents its knowledge in an lexically abstract fashion will it be able to generalize the use of a word from one context to another. Hadley is not very explicit about what he means by "syntactic position" in this definition. Notions like grammatical subject or object are not part of the training data itself, but are categories imposed by the theorist. Thus, different theorists may differ in terms of what constitutes a novel syntactic position. For instance, is the direct object of a sentence whose subject is a proper name a different syntactic postion from the direct object of a sentence whose subject is a noun phrase of the form Determiner-Adjective-Noun? Our notion of structural abstraction is intended to get at precisely this question. We can imagine a variety of degrees of structural abstraction, depending on the class of sequence types that are treated in an analogous fashion by the network. In our experiments, we will explicitly address this topic in the context of anaphoric reference.

2. SURVEYING THE GRAMMAR OF ANAPHORA

Before we turn to a discussion of our network experiments, it will be useful to lay out the empirical landscape that we will explore. Cross-linguistically, it is a basic fact that reflexives like *herself* and *himself* show restrictions on the noun phrases from which they take their interpretation, their *antecedents*. For instance, the reflexive direct object *herself* in (2) may take as its antecedent *Mary's mother*, but not *Mary*.

(2) Mary's mother admired herself in the mirror.

Similarly, the reflexive *himself* in (3) may have as its antecedent only the noun phrase *the man who knows John*, and not *John*.

(3) I asked the man who knows John about himself.

Such restrictions have been characterized in terms of the structural relation between nodes in a syntactic structure called *c-command*, which can be defined for present purposes as follows:

(4) X c-commands Y if the parent of X dominates Y and X does not dominate Y.

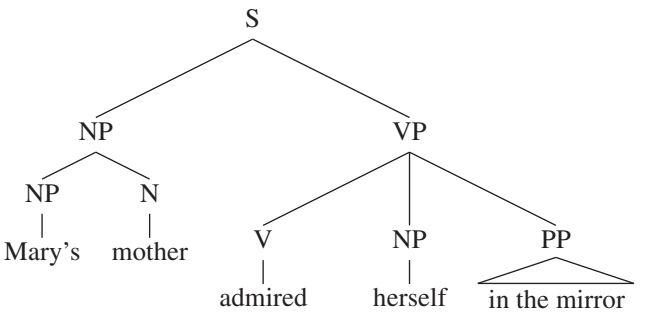
Using this relation, we can define a condition on English reflexive interpretation:

(5) A reflexive must have as its antecedent a c-commanding noun phrase.

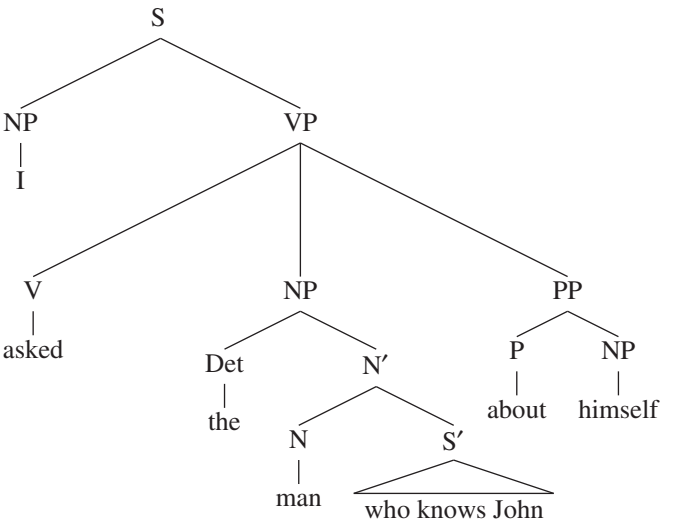
[4]Hadley also defines a notion of *weak systematicity*, requiring only that the an agent be able to process sentences that are novel combinations of its vocabulary, without the generalization of words to novel positions. One might argue that a network satisfying this condition demonstrates an extremely limited sort of lexical abstraction, but no more than the sort that might be exhibited by a bigram model, which can process novel sequences so long as the constituent bigrams have been observed during training.

If we look at the standardly assumed structural descriptions for the sentences in (2) and (3), as shown (6) and (7) respectively, we see that the possible antecedents for the reflexives all stand in a relation of c-command: In the first case, the parent of the NP *Mary's mother* is the S, which dominates the reflexive, and in the second case, the parent of the NP *the man who knows John* is VP, which dominates the reflexive.

(6)



(7)



In contrast, none of the impossible antecedents c-commands the reflexive, as the reader can readily confirm. The c-command condition alone is not sufficient to characterize the possible antecedents for a reflexive. For instance, in (8), *herself* can refer only to Alice in spite of the fact that the NP *Sue* c-commands it as well.

(8) Sue thinks that Alice mentioned herself.

There is an additional requirement, then, that reflexives be "close" to their antecedents, which we can formulate very roughly as follows:

(9) A reflexive must have as its antecedent a c-commanding noun phrase in the same clause.

Pronouns like *him* and *her* show an interestingly different generalization. Consider, for instance, examples like those just discussed but with the reflexives replaced by the pronouns:

(10) Mary's mother admired her in the mirror.

(11) I asked the man who knows John about him.

In these cases, the pronoun's interpretation is restricted in a way that is just the opposite of that seen for reflexives. Namely, the pronoun *her* in (10) can refer to any female individual except Mary's mother (even someone not mentioned in the sentence), and the pronoun *him* in (11) can refer to any male individual except the man who knows John. This pattern points to the following constraint on English pronoun interpretation.

(12) A pronoun may not take as its antecedent a c-commanding noun phrase.

In fact, this condition is a bit too strong. In the following sentence, pronoun *her* can in fact take the c-commanding name *Sue* as its antecedent.

(13) Sue said that John had visited her.

The contrast between this case and those just discussed suggests the existence of a locality restriction for the rule of pronoun interpretation, similar to the one we observed for reflexive interpretation. Roughly speaking, pronouns in English may take as their antecedents c-commanding NPs, so long as the NP is outside of the pronoun's clause. We can therefore revise our condition on pronoun interpretation as follows:

(14) A pronoun may not take as its antecedent a c-commanding noun phrase within its own clause.

Note that it is not our purpose here to characterize in a more precise way the notion of locality relevant for pronoun and reflexive interpretation, and the experiments we describe will not probe this aspect of anaphoric interpretation.[5]

In the next section, we turn to experiments that attempt to model the acquisition of these anaphoric dependencies. Before doing that, it will be useful to address one sort of objection that could be raised at this point. The constraints on anaphoric interpretation given in (9) and (14) are similar to those one finds in linguistics texts and are usually taken to represent part of the abstract system of linguistic knowledge a speaker possesses. Such a model of linguistic competence is to be distinguished from a model of linguistic performance, which would specify the mechanisms that underlie online processing. Connectionist modeling typically eschews drawing a distinction between competence and performance. Stated less contentiously, it aims to directly model human processing. Accordingly, one might argue that these abstract constraints on pronoun and reflexive

[5]Indeed, languages show a certain degree of variation in what characterizes the local domain both for pronouns and reflexives, though interestingly there is almost always complementarity between the contexts in which pronouns and reflexives may occur with a particular antecedent. For some relevant discussion, see Koster & Reuland (1991). We put aside discussion of this issue for the remainder of this article, apart from noting that in current work we are exploring the question of whether there are any properties of SRNs that would lead us to expect such complementarity.

interpretation are simply irrelevant. Instead, on this view one should be studying and directly characterizing patterns of human sentence processing.

There are two lines of response to this objection. First, it is important to note that the patterns of acceptability that are modeled by abstract contraints on anaphoric interpretation are not simply a linguist's fantasy. When such patterns are subjected to careful scrutiny using carefully controlled experimental methodologies and statistical analysis, the results are a virtually perfect match for the judgments reported in the theoretical linguistics literature (Sprouse & Almeida 2012). Consequently, any theory of human linguistic behavior must provide an account of the contrasts that theories of linguistic competence attempt to model.

Secondly, whether or not one rejects the notion of linguistic competence and the importance of modeling it, with respect to the anaphoric phenomena under discussion, the empirical patterns of coreference possibilities that are found in studies of sentence processing are precisely aligned with those that are predicted by the abstract conditions on coreference given previously. Experimental results from Nicol & Swinney (1989), Gordon & Hendrick (1997), Asudeh & Keller (2001), Badecker & Straub (2002), and Sturt (2003) uniformly confirm people's sensitivity to a c-command–based locality condition in pronoun and reflexive interpretation, along the lines of those in (9) and (14). Of course, abstract structural constraints like (9) and (14) do not inform us about the time course in which interpretation will take place, nor do they tell us the way in which properties of the discourse will affect the interpretation process. However, given that they conform to the output of online sentence processing, it strikes us as a reasonable first step to try to model the patterns of interpretation they generate.

It is nonetheless important to point out that some of the experimental results just cited as well as those from Runner, Sussman & Tanenhaus (2003) do find limited divergences between the results of online processing and those predicted by abstract grammatical principles like (9) and (14). Note, however, that the context in which a divergence between competence and performance has been argued to exist, the so-called *picture*-NP construction, is notorious even within the theoretical syntax literature. For instance, in this domain pronouns and reflexives no longer exhibit their usual complementarity of interpretive possibilities.

(15) Joe saw pictures of himself/him in the newspaper.

This property and others have led to a number of innovations in the characterization of the grammatical conditions on coreference (Chomsky 1986; Pollard & Sag 1992; Reinhart & Reuland 1993), which remain the topic of considerable debate. Thus, it would be overstating the empirical situation rather substantially to say that these experimental results falsify a grammatical account of anaphoric interpretation. Because we focus our attention in this article on core cases of anaphoric dependencies within a single clause, any minimal differences among the different grammatical proposals or possible divergences between competence and online processing that reside outside this limited domain are simply not relevant to our explorations. An anonymous reviewer suggests that the very existence of these noncore cases might be taken as evidence against a rule-based approach to anaphoric dependencies, instead motivating an emergentist approach in which quasi-regular generalizations (Plaut et al. 1996) on anaphoric dependencies are induced from the available data. Even if correct, such an account must still explain the sharp generalizations in the core cases. We take our present experiments to be an attempt to determine the feasibility of precisely such an approach. Moreover, the mere existence of exceptions does not necessarily argue against grammatical accounts of our knowledge: a number of recent proposals

deal with exceptionality through the use of violable constraints, whose satisfaction depends on their interaction with other constraints on form and interpretation (Fischer 2004; Wilson 2001).

3. EXPERIMENT 1: ESTABLISHING ANAPHORIC REFERENCE VIA WORD PREDICTION

Our first experiment explores the ability of SRNs to learn to assign an interpretation for reflexives and pronouns, in accordance with the constraints discussed in the previous section. This task differs in a crucial respect from those to which SRNs have been applied previously. In Elman (1993) for instance, training and testing consisted entirely of word prediction: The network was trained to maximize its likelihood of predicting the next word, and the network's knowledge of grammatical structure was assessed in terms of its success at prediction in certain contexts. For instance, by looking at the class of verbs predicted by the trained network immediately after the input *The man who they like* and other similar contexts, we have some indication as to whether it demonstrates sensitivity to a structurally conditioned constraint on subject-verb agreement. Knowledge of anaphoric interpretation, in contrast, cannot be completely assessed via word prediction. In this task, the grammatical knowledge that the network acquires during training must be put to use not only in predicting the next word (for example, to predict that *herself* but not *himself* is a possible next word after the sequence *Mary who John likes saw*), but also in activating a semantic representation of the individual to which the reflexives and pronouns may refer. Therefore, it is not sufficient to train and test the network on the word prediction task alone.

Nonetheless, one of the claims of SRN-based language work is that there is something fundamental and cognitively natural about the prediction task in language learning. As Elman (1991) argues, this task avoids the necessity of an external teacher, as the target output of the network is provided by the environment at the next moment in time. Elman also notes that the presence of these target outputs provides a (partial) solution to the problem of overgeneralization, since a learner's faulty predictions may lead to conflict with subsequent input. Finally, Elman points to data from human performance that demonstrates that human listeners are indeed adept at making predictions for subsequent words in a sentence and that they show distinctive brain responses (N400) in the face of violations of their expectations.

Because the prediction task has played a fundamental role in SRN models of grammar, and because past successes in word prediction appear to point to the conclusion that SRNs can induce hierarchically structured grammatical representations, we decided to explore the following question in conjunction with the two listed earlier:

3. To what degree do the representations induced by an SRN trained to perform word prediction support the computations underlying anaphoric interpretation?

To address this issue, we separated the training of the network into two phases. In the first phase, we trained on the word prediction task an SRN structured like the one used by Elman (1993), with the following components:

- An input layer of 28 units was used to encode the identity of the current word (or a sentence-boundary token) in a localist fashion.
- An output layer of 28 units represented the predicted next word.

- A pair of hidden layers of 10 units each were placed immediately after the input layer and before the output layer. These layers allowed the network to reencode the localist input and output representations in a distributed fashion. For expository purposes, we will refer to these as (re-)coding units.
- A recurrent hidden unit layer of 70 units was placed between the pair of layers of recoding units. At each time step in processing, the activation levels of these units were copied to a set of 70 additional "context'' units. At the next time step, the activation levels of the context units were provided as input to the original 70 hidden units through connections

The resulting architecture is depicted in Figure 1.

This network contained sigmoidal hidden units and normalized exponential output units, the latter allowing us to interpret the activation values at the outputs as the network's predicted probability distribution for the next word. We assume, following previous work, that this first phase of training on the word-prediction task will suffice to allow the network to establish representations of the sentences in the training corpus that encode the hierarchical relationships necessary to resolve, for instance, subject-verb agreement. If this representation also has the properties that

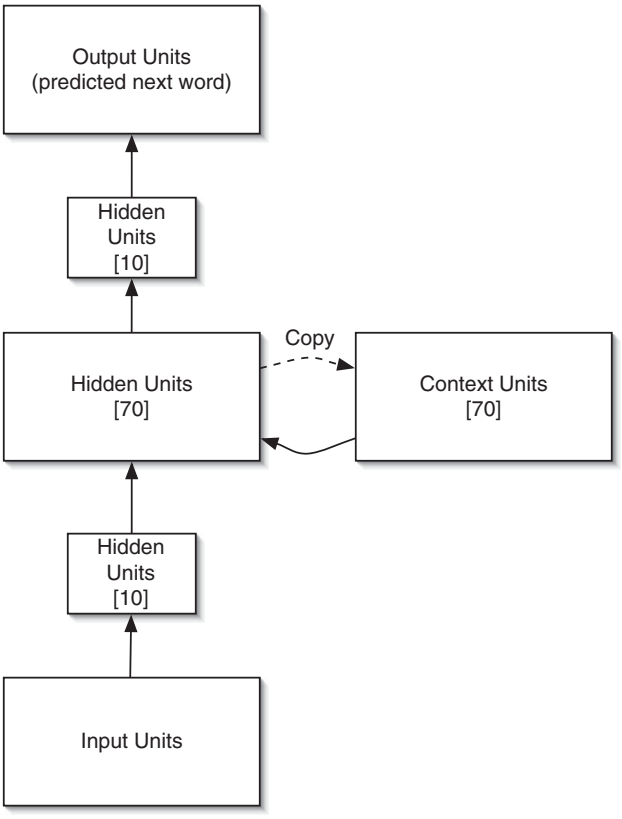


FIGURE 1 Phase 1 word prediction network.

linguists have found necessary for characterizing grammatical phenomena such as constraints on anaphoric interpretation, then it ought to suffice as input to a network that assigns interpretation. (See also Cleeremans, Timmermans & Pasquali 2007 for a similar training regimen, though in a rather different domain.) Following this reasoning, the weights of the Phase 1 network were frozen prior to the initiation of the second training phase. Then, the activation values from the 70 hidden units of the Phase 1 network were used as inputs to a new, interpretive network. We used an SRN for this interpretive network in order to give this architecture the best possible chance of success. The SRN provided a sort of "memory'' for past states of the word-prediction network. Without such an ability, the network might not be able to resolve sentence-final pronouns, which may be assigned trivial representations in the prediction network since sentence-final words are of no predictive value. The architecture of this interpretive network was as in the prediction network, with the difference that there was no new input layer, and the localist output layer contained a unit for each distinct referent (see Figure 2).

This interpretive network was then trained to assign an "interpretation" to each word in the sentence, using as input the hidden units of the Phase 1 network. What we mean by *interpretation* is rather simple: Each word in our language, other than reflexives and pronouns, was associated with a single output unit, which we take to constitute a localist representation of its interpretation. Thus, when the word *John* is presented as input, this should trigger the activation of the "John'' interpretive output unit, the verb *admires* should trigger the activation of the *admires* interpretive

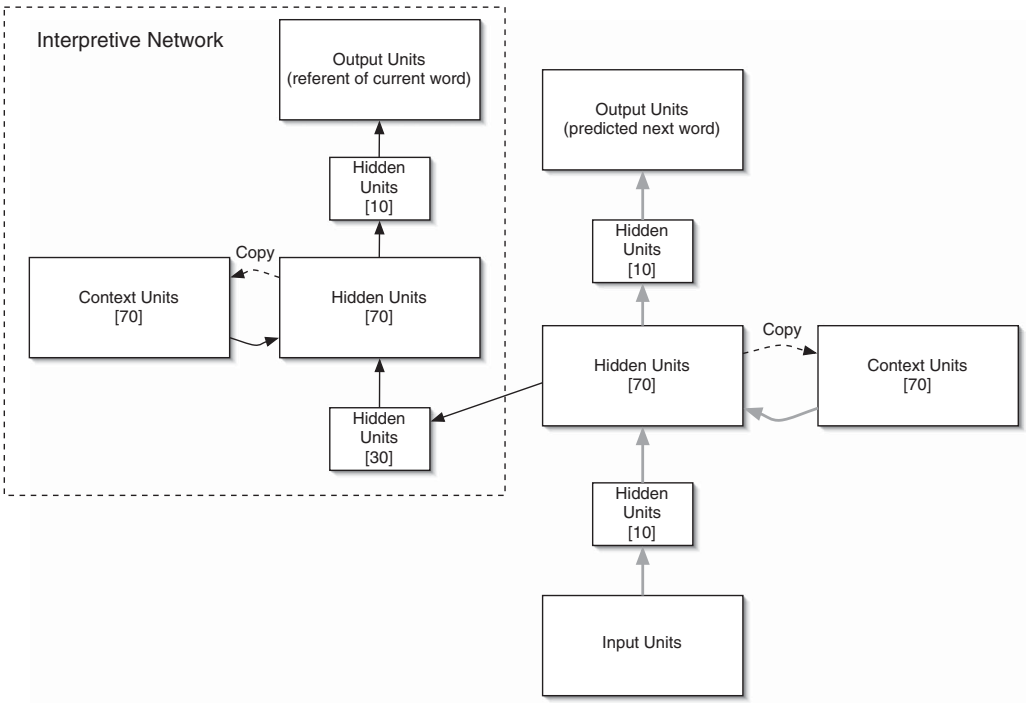


FIGURE 2 Phase 2 network (word prediction + interpretive network).

output unit, and so on. The interpretation associated with reflexives and pronouns, in contrast, depends on the context in which these words occurred. In a sentence like *John admires himself*, the reflexive would be associated with the interpretative output associated with the name *John*, since that is the only grammatically licensed possibility, while in the sentence *Nate admires himself*, the same word *himself* would be associated with the interpretative output assigned to *Nate*. In the artificial language we presented to the network, this choice was always unambiguous: The structures of the language in combination with the hierarchically based constraint in (9) always determined a unique possible antecedent for the reflexive. For pronouns, this was not the case. The constraint in (14) tells us what may not function as the antecedent of the pronoun, but does not say what must. Thus in general the choice of referent for a pronoun cannot be made deterministically. For a sentence like *John admires her* then, the pronoun might have as its referent any of the female individuals in our domain; likewise for *John admires him*, the pronoun may refer to any male individual in the domain except John. For each occurrence of a pronoun in our training data, a particular gender-appropriate referent was randomly selected among those that were compatible with the restriction in (14), and the network was trained with this referent as the interpretive target in this case.[6] It is worth keeping in mind this nondeterminism of the pronominal reference task, as we might expect it to lead to the network having more difficulty with pronoun than reflexive interpretation.

3.1. Corpus Design and Training Regimen

For training and testing, we focused our attention on an artificial language modeled on a subset of English. As in Elman's original work, this language consisted of simple (transitive) sentences showing subject-verb agreement and including optional modification of noun phrases by relative clauses. In this language, noun phrases were typically of one of three types: a proper name (optionally modified by a relative clause), a pronoun, or a reflexive. Because all of our noun phrases were singular, we used gender as the feature triggering verb agreement. Note also that we permitted pronouns and reflexives to occur only in object position and that the gender of the reflexive was required to match the gender of the antecedent. The language, then, included sentences of the following forms:

(16) a. Subject verb agreement:
John sees-M Mary; Mary likes-F Bill
b. Subject and object relative clauses:
John who sees-M Sue admires-M Bill; Alice likes-F Nate who Mary admires-F
c. Reflexive and pronominal objects:
John sees-M himself; Alice who Harold likes-M admires-F her

[6] An anonymous reviewer observes that it might be more reasonable to provide probability distributions as target outputs. We do not do this, however. We tentatively assume that these targets stem from the learner's hypothesis about the interpretation of a pronoun. Such hypotheses are derived from multiple clues, including aspects of the discourse context as well as her understanding of the speaker's intentions. Even if the learner's hypothesis is unlikely to be correct in every occasion, there is little reason to expect that this hypothesis will match the set of structurally accessible antecedents. Nonetheless, even in the face of unique target outputs, the optimal output (in the sense of minimizing error) will correspond to a probability distribution over possible outputs. In future work, we plan to explore the effect of introducing noise into the targets for pronominal reference.

Our artificial language also included one additional type of noun phrase: For each of the six names in our fragment, we introduced a new noun that could occur only as the object of a sentence when its corresponding name was the subject of that sentence. Thus, the noun *junipers* occurs only as the object of sentences with subject *John* (possibly modified by a relative clause), the noun *mangos* occurs only as the object of sentences with subject *Mary*, and so on. These sentences can be thought of as the linguistic expression of distinctive semantic properties for each of the names in our domain. Consequently, the resulting corpus used for training incorporates semantic restrictions on word co-occurrence (cf. Rohde & Plaut 1999) of a sort found in natural text, though our restrictions are admittedly rather coarsely grained. This addition to our language also has the effect of requiring the network to represent different names distinctly in order to carry out word prediction, as the identity of the name in subject position will determine possible unique object noun phrases. Without these unique objects, the word prediction network would have no incentive to represent the names of a given gender distinctly from one another, as the identity of the name (within a gender) has no predictive value for determining the next word. And without such distinctive representations, it would be impossible for the interpretive network to recover the grammatically accessible referent(s) for the pronouns and reflexive. Indeed, the requirement that the identity of words be kept distinct is part of what distinguishes Hadley's (1994, 2003) notion of *Strong Semantic Systematicity* from the notion of *Strong Systematicity* defined in (1) (the strong version requiring the ability to assign meanings to novel combinations of words) and is the key component of Marcus's (1998, 2001) conception of universally quantified one-to-one mappings. Networks trained to perform prediction will tend to reduce error by collapsing their representations of words with similar distributions, thereby leading to generalization. However, it is less clear what will happen when the words must be kept distinct in order to satisfy other task demands.

The fragment described in the previous paragraph can be characterized using the context-free grammar shown in Figure 3.

Using the Simple Language Generator tool (Rohde 1999b), constraints were added to this context-free backbone to enforce subject-verb agreement, subject-reflexive agreement, and subject-distinctive object agreement. Furthermore, probabilities were associated with each rule to produce a stochastic grammar. Specifically, the probability of a complex NP (i.e., one containing a relative clause) was approximately 25%, equally split among the two types of relatives. Among direct objects, the probability of a reflexive was 12.5%, as was the probability of a direct

S	→	NP VP
NP	→	Name (Rel)
VP	→	V NP-obj
NP-obj	→	Name (Rel) | Refl | Pronoun | Distinctive-Obj
Rel	→	who VP | who NP V
Name	→	John | Harold | Nate | Mary | Alice | Sue
Refl	→	himself | herself
Pronoun	→	him | her
Distinctive-Obj	→	junipers | hotdogs | nachos | mangos | avocados | salamanders
V	→	sees-M | loves-M | admires-M | kisses-M | visits-M | sees-F | loves-F | admires-F | kisses-F | visits-F

FIGURE 3 Grammar for the training corpus.

object pronoun. We then stochastically generated a training corpus of 32,000 sentences along with a (potentially overlapping) test set of 5,000 sentences from this language, the latter used to assess generalization performance during training. These sentences had a minimum length of 3 words and were truncated to a maximum length of 15 words (average length was 6.2 words). Of these, approximately half (49%) were complex in the sense of including at least one relative clause.

Readers familiar with the results of Elman (1993) may be concerned about this high percentage of complex sentences. Elman reported that his SRN was able to acquire the structural regularities of subject-verb agreement only when the initial training data contained no complex sentences, with complex sentences being introduced only at later stages in training. Indeed, in Elman's simulations, the training data reached 50% complex sentences only during the second half of training, after the network had already established representations capable of carrying out word prediction for simple sentences. However, Rohde & Plaut (1999) have demonstrated that this kind of staged training regimen, which "starts small," yields better learning starting with a more complex training corpus only when the network's initial weights are constrained to be very small, as in Elman's simulations where the initial weights were in the range $[-0.001,+0.001]$. In the simulations on which we report, we follow Rohde and Plaut in taking the initial weight range to be $[-1,+1]$, thereby eliminating the need to stage the training.[7]

Both Phase 1 and Phase 2 networks were trained using backpropagation through time (BPTT) with a learning rate of .001 and no momentum. Training was done in batches of 51 sentences (i.e., weight updates were done on the basis of the error gradient computed through this set of sentences). This set of sentences was large enough to provide reasonable coverage of the grammatical domain and permitted a fairly accurate assessment of the true error gradient. Smaller batches yielded less successful learning, and larger ones did not improve accuracy. The SRN model, as originally studied by Elman, made use of a different learning algorithm from the one we are using, namely standard backpropagation. We are, however, interested in understanding the range of generalizations that the SRN architecture can detect. In recurrent networks, simple backpropogation only computes an approximation to the error gradient, unlike BPTT, which computes the gradient exactly. Consequently, during simple backpropagation learning it is possible that weight updates will increase error on the examples on which it was just trained rather than decrease it. While it is possible that such failure to follow the error gradient will give rise to a kind of inductive bias that will favor the learning of linguistic generalizations, experiments we have run with simple backpropogation learning have yielded results that are significantly worse than the results obtained with BPTT. In the studies we report here, we therefore employ BPTT, and follow Rodríguez (2001) in continuing to call this an SRN architecture. Batches of sentences were chosen for training by sampling uniformly from the fixed training set, with "online'' updates made to the weights using a cross-entropy cost function. Training was stopped when error on the test sample stopped decreasing, which occurred after 160,000 updates.

[7]For a number of the simulations reported here, we have also performed identical simulations using Elman's starting small regimen, in which the network's hidden units' activations were reset during training after an increasingly long delay. In Elman's paper, this had an effect on learning identical to that of changing the relative proportion of simple and complex sentences over training, namely it permitted the network to learn the long-distance subject-verb dependencies. In none of our simulations did we find that this starting-small regimen improved the network's performance, either in terms of error minimization or in terms of generalization to held-out sentence types.

3.2. Results

We trained six networks as described above, each with different random initial weights. We quantitatively evaluated the performance of these networks in a number of ways.

3.2.1. Word Prediction Accuracy

Because any initial portion of a sentence may be completed by a variety of words, it is unfairly strict to rate the network's prediction as incorrect when it does not uniquely pick out the actual word in the sentence from the test set on which it is being evaluated. Instead, we exploited the fact that we used a stochastic grammar to generate our training and test data and compared the true probability distribution for next words as determined by the probabilities on the grammar rules to the distribution we could read off of the levels of activation of the output units. (Recall that the output units are normalized.) Following Rohde & Plaut (1999), we assessed the difference between these two distributions in terms of Kullback-Leibler divergence, a measure of the entropy of one distribution (in our case, the target) given another (the network's hypothesis). Kullback-Leibler divergence is defined as follows (where t is the target probability distribution as determined by the grammar and o is the network's predicted distribution):

$$D(t||o) = \sum_i t_i \log \frac{t_i}{o_i}$$

The mean divergence error per word prediction on an out-of-sample test set of 20,000 sentences was .029. While this number is of little interest on its own, it is worth noting that it is very slightly worse than the results reported by Rohde & Plaut (1999) using a similar training regimen and SRN architecture. However, it should be noted that the addition of the distinctive objects makes the task of word prediction considerably more difficult. Note furthermore that because this test set is disjointed from the stochastically generated training set, it contains only complex sentences involving at least a single relative clause, as the simple sentences were all contained in the training set.

3.2.2. Agreement Accuracy

To ensure that our Phase 1 network was succeeding in the subject-verb agreement domain that has been the focus of much previous work, we tested the accuracy of the network's predictions for verbal agreement. An error was scored if a single output unit representing a nonagreeing verb was more active than any of the units representing an agreeing verb. Put another way, the network is correct in predicting a feminine verb only if all feminine verb units are more active than all masculine units, and vice versa if the target was a masculine verb. Even under this rather strict measure, the network's performance on the agreement task was outstanding. For six networks with different random initial weights, the average error rate on agreement was 0.2% over the same set of out-of-sample (complex) test sentences.

3.2.3. Distinctive Object Predictions

We evaluated in further detail one aspect of the network's performance on prediction, specifically that concerning the prediction of the distinctive objects whose distribution is determined by the subject. We did this by considering the predictions of the network at each main verb on the out-of-sample test set used to assess word-prediction accuracy. The network was considered to have performed correctly if the network's activation of the appropriate distinctive object was higher than all others by a margin of at least .05. Over six networks with different random initial weights, the average accuracy on distinctive object prediction was 97.2%.

3.2.4. Anaphora Accuracy

We measured the accuracy with which the interpretive network assigned the correct referent for reflexives. An error was scored if the correct referent did not receive the highest activation. Under this measure, our six trained networks achieved an average accuracy of 89.7 $\pm$ .6% over a set of novel (complex) sentences.[8,9] Quantitative assessment of accuracy on pronominal reference is somewhat more difficult because there is no deterministic solution to the problem, i.e., multiple interpretations for the pronoun may be grammatically licensed. In order to count as an accurate interpretation, we required that all of the grammatically possible interpretations be active to at least a level of .1, and that the lowest among these be more active than the highest impossible antecedent by a margin of .05. With this somewhat lenient criterion, our six trained networks achieved average accuracy of 71.9 $\pm$ 1.5% for the interpretation of pronominal objects for a set of sentences whose subjects were modified by a single relative clause. For sentences with subjects including two levels of embedding accuracy on pronominal objects was 62.2 $\pm$ 2.6%. These results suggest on first blush that the network has been reasonably successful at solving the reference task, particularly in the case of reflexives, and that the hidden unit representation derived through word prediction is indeed structurally rich enough to support interpretive processes.

3.3. Linearity Effects

However impressive the quantitative results just discussed are, it is also important to note that the network's behavior in the anaphoric interpretation task diverged in certain systematic ways from what we would expect from English speakers. Consider, for instance, the performance of one of our networks on the sentences in (17).

[8]Here and throughout, percentages reported in this way give a 95% confidence interval.

[9]This measure of accuracy on reflexive interpretation might be seen as too lenient: Since network activations have a natural interpretation as probabilities, a referent having the highest activation is only the one that the network takes to be most likely. Given that the interpretation of reflexives is always unique given a particular structure in our training set, it might be appropriate to require a nearly unimodal probability distribution to count as success. One way to test this is to require that the activation of the target referent must be above some threshold. If we set this threshold to be .8, average network accuracy decreases slightly, to 77%.

(17) a. Alice who Mary loves admires herself
herself: $p(Alice) = .99$
b. Alice who loves Mary admires herself
herself: $p(Alice) = .80$, $p(Mary) = .16$, $p(Sue) = .03$

For both of these examples, judgments and online processing data (Xiang, Dillon & Phillips 2009) of English speakers point to the unavailability of *Mary* and *Sue* as antecedents for the reflexive *herself*, in accordance with the grammatical constraint in (9) since both fail to locally c-commands *herself*. As indicated by the probabilities listed below example (17a), the network correctly interprets this case, assigning high probability to *Alice* as the unique possible antecedents of *herself*. But for (17b), which has the same structural relation between *Alice* and *herself*, the network assigns significant probability mass to *Mary*, incorrectly, since it is not the subject of the sentence of which *herself* is the object.

From the perspective of the linguistic analyses of anaphora reviewed in section 2, this contrast is puzzling in that the examples do not differ one from the other in any structurally relevant way. Indeed, the difference in the network's performance on these sentences appears to be tied to a linear rather than structural factor: the presence in (17b) of a linear sequence that forms a possible simple sentence, *Mary admires herself*. In this sequence, if it were taken to be an independent sentence, *Mary* would be a possible antecedent for *herself*, and we suggest that the network is basing its response on the union of the antecedents that would be possible structurally and linearly. In example (17a), there is no such linear sequence, and as a result the network correctly bases its decision on structural factors. Example (18) shows that the linearity effect can also dissolve when the "linear subject'' is not a possible subject in virtue of its case morphology.

(18) Alice who loves her admires herself
herself: $p(Alice) = .99$

Because the sequence *her admires himself* is not a possible sentence, the network appears to rely upon a structurally based strategy to (correctly) interpret the reflexive.

To test whether such linearity effects are in fact systematically observable, we analyzed the behavior of a number of networks on larger classes of sentences. We focused our initial analysis on sentences involving reflexives, since our networks achieved higher levels of performance for these anaphoric elements.[10] We specifically compared the networks' performance on sentences whose subjects were modified by a relative clause (possibly subject or object) containing a transitive verb. We required that the name in the subject and the one within the relative clause be of the same gender, which was in turn identical to the gender of the object reflexive. We generated all of the sentences of this form that are permitted by the grammar up to two levels of embedding,

[10]We conducted a similar linearity analysis on sentences containing pronominal objects, using the comparisons described in section 5.2. Though this analysis did turn up many instances of linearity effects across all six of our networks, there were also three instances of statistically significant effects in the direction opposite from what is predicted by linearity. As already noted, however, the overall level of performance in the task of pronominal reference was not very high, and network outputs did not closely reflect the desired activations: When the target activation was 0, the mean activation across networks was .292, when the target was .333 mean activation was .287, and when the target was .5, mean activation was .32. As a result, we are reluctant to conclude much from these results.

TABLE 1
Mean Activation of Linear Distractor as Reflexive Interpretation (Correct Activation = 0)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.028	.061	361.7	<.001
Net B	.071	.151	736.4	<.001
Net C	.048	.171	2243	<.001
Net D	.053	.100	541.6	<.001
Net E	.084	.082	0.25	n.s.
Net F	.049	.056	13.18	<.001

TABLE 2
Mean Activation of Subject Referent as Reflexive Interpretation (Correct Activation = 1)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.887	.824	340.3	<.001
Net B	.822	.695	845.7	<.001
Net C	.817	.641	1572	<.001
Net D	.792	.621	1403	<.001
Net E	.740	.557	1181	<.001
Net F	.849	.581	3617	<.001

including both subject relative (17b) and object relatives (17a), a set which numbered 9,300.[11] We reasoned that if the network is sensitive to effects of linearity, there should be detectable differences in activations of potential referents for the reflexive between subject and object relatives, as only subject relatives give rise to the appropriate linear sequence. Specifically, if a name *N* that is linearly adjacent to the main verb is in some sense taken to be the subject of the sentence, it should induce a coreference effect with the reflexive object, resulting in higher activation for *N*'s referent as a possible interpretation of the reflexive. We tested this prediction on six separate networks with different random initial weights, each trained as discussed in the previous section. This result was highly significant in the predicted direction in all but one of the networks. The activation means, and *F*-ratios and *p* values are reported in Table 1.

We also compared the activation of the subject's referent in these cases, reasoning that the linear distractor present in subject relatives is more likely to draw away activation from the target antecedent than the nonlinear distractor present in object relatives. This result is highly significant in all of the networks, as shown in Table 2.

[11] When a sentence included two levels of relative clause embedding, the outer relative was always a subject relative. The gender of the object of the outer relative was allowed to vary freely. The two sentence types are thus illustrated by the following pair of sentences:

(i) a. ObjRel: John who sees Alice who Harold visits admires himself.
b. SubjRel: John who sees Alice who visits Harold admires himself.

The consistent differences that we find between subject and object relatives suggests that we are not simply seeing the effect of a noun intervening between the subject and the main verb, but rather a noun that stands in a particular relationship (linear adjacency) with the verb.

We do not take the presence of these linearity effects to show that the phase one network's hidden units provide an insufficiently rich representation of the syntactic structure to support the task of structurally determined anaphoric reference. On the contrary, the response patterns we observe can be interpreted as demonstrating a sensitivity to syntactic structure: In (17b), for instance, we assume that *Alice* is taken to be a possible antecedent of the reflexive (in contrast to *Sue*) precisely because it is the subject of the main clause and therefore c-commands the reflexive. Similarly, mean activations for the (correct) subject referent for sentences involving subject relatives, found in Table 2, remain well above .5. The presence of this pattern, in which the subject remains active, together with the consistent difference between subject and object relatives in our statistical tests, undermines an alternative potential explanation for the difficulty of such cases for the network, which is based on the difficulty of maintaining activation across intervening material. Nonetheless, the presence of these linearity effects does point to a different sort of deficiency: The networks are sensitive to too rich a set of possible contingencies. In establishing anaphoric dependencies, the network was apparently unable to ignore a probabilistically, but not categorically, reliable generalization about the importance of the identity of a name in name-verb-reflexive sequences that do not form sentences hierarchically.

One might imagine that the network's behavior is the result of properties of the training data to which it is exposed. As a statistical learner, the backprop-trained SRN is rewarded for exploiting any regularities in the training data that it can. If a reflexive occurring in a N-V-reflexive sequence is reliably interpreted as coreferent with the N, attention to this property could help the network to reduce error in anaphor interpretation. In order to test whether our training data were misleading in this respect, we compared the reliability of linearity as a cue to anaphor interpretation in naturally occurring text and speech data with our training data. The reliability of the linearity cue was assessed by computing the probability of coreference between a simple noun phrase and a reflexive, given that the noun phrase immediately precedes a verb phrase in which the reflexive is an argument. That is,

$$p(\mathrm{NP}_i\ \mathrm{V}\ \mathrm{refl}_i \mid \mathrm{NP}\ \mathrm{V}\ \mathrm{refl}) = \frac{C(\mathrm{NP}_i\ \mathrm{V}\ \mathrm{refl}_i)}{C(\mathrm{NP}\ \mathrm{V}\ \mathrm{refl})}$$

We calculated this probability for third person reflexives in our training data as well as in three (parsed) corpora: the Penn Treebank (Marcus, Santorini & Macinkiewicz 1993) versions of the Brown Corpus (Francis 1964), the Switchboard Corpus (Godfrey, Holliman & McDaniel 1992), and the *Wall Street Journal* corpus. Using the tgrep2 tool, we extracted all sentences with a third person reflexive (either *himself*, *herself*, or *themselves*) following a verb, and from this set of sentences we handselected only those in which some nominal element was separated from this verb by only auxiliary verbs or adverbial heads.[12] In addition, we extracted all instances of reflexives in child-directed speech in the U.S. English portion of the CHILDES database (MacWhinney &

[12]This puts aside a good many occurrences of reflexive pronouns, all inconsistent with a linearity-based generalization. However, these other sentences do not bear on the question of whether linearity is a possible source of information in the context in which it is confounded with a structural relation. Furthermore, as we shall see, SRNs appear to learn

TABLE 3
Linearity as a Cue to Coreference

Corpus	C(NP_i *V refl*$_i$)	C(*NP V refl*)	$p(NP_i\ V\ refl_i \mid NP\ V\ refl)$	*95% interval*
Brown	127	144	.88	[.82, .93]
Switchboard	30	31	.97	[.84, .99]
WSJ	45	62	.73	[.60, .82]
CHILDES	73	75	.97	[.90, .99]
Training data	4531	5005	.91	[.90, .91]

Snow 1985) and counted the occurrences of the same kinds of sequences, along with number of times that the linearly adjacent nominal was the antecedent of the reflexive. The counts are presented in Table 3, along with confidence intervals for the estimates.

With the exception of the *Wall Street Journal*, none of these corpora has a conditional probability that is significantly different from that of our training data. The lower conditional probability present in the *Wall Street Journal* corpus would be useful to dissuade a statistical learner from adopting a linearity hypothesis. However, the lower conditional probability arises because of the greater complexity of subject NPs in the *Wall Street Journal* corpus as compared to what is found in spontaneous speech. Indeed, in the spoken corpora, we found extremely few examples of sentences whose interpretations were not consistent with a linearity-based generalization. In Switchboard, there was only one such example (“and most kids these days have gotten themselves into, uh, financial situations, where they have to be working all the time’’), while in the CHILDES data there were two, both of which seemed to be speech errors (“It is washing herself in high water” and “you go herself”).

As far as we are aware, there is no evidence for the use of linear-based generalizations in the human processing of anaphora. We hypothesize that this distinction between human and network performance is due to the lack of appropriate inductive bias in the network to ignore spurious linear generalizations in extracting grammatical generalizations.

Whether there are effects of linearity in other aspects of human sentence comprehension comparable to what we observed in our network model remains a matter of debate. For example, in the case of subject-verb agreement, it has been suggested that interference from a local noun like *cabinet* in (19) is evidence for the influence of linearity over hierarchy (Hemforth & Konieczny 2003).

(19) The key to the cabinets were missing.

However, it is unclear that the relevant property of the false “attractor’’ of agreement is actually linear adjacency to the agreeing form: Comprehension studies find substantially less interference from adjacent local nouns in clausal modifiers like (20) than in phrasal mophrase-modifiers like (19) (Pearlmutter, Gakasey & Bock 1999; Nicol, Foster & Veres 1997).

(20) The key that opened the cabinets were missing.

distinct generalizations about anaphoric dependencies in distinct structural contexts. Consequently, it seems appropriate to examine the support for such a generalization in a particular context.

This suggests that the hierarchical role/position of the local attractor in the subject noun phrase, and not its linear proximity to the verb, accounts for its power to interfere. Instances in which linearity does appear to affect agreement (in language production), such as those observed in the context of disjunctive coordination (Haskell & MacDonald 2005), are arguably cases where no hierarchical property distinguishes the candidate agreement sources. However, "local coherence'' effects on other components of the comprehension process (Tabor, Galantucci & Richardson 2004; see also Duffy, Menders & Morris 1989, and Konieczny 2005), as well as less extreme departures from hierarchy (e.g., "Good Enough Parsing'' effects observed in Christianson et al. 2001 and Ferreira, Christianson & Mollingworth 2001) do suggest that some local parsing alternatives are activated, even if only briefly, despite their global incompatibility with the preceding syntactic context. If, in fact, mere linear adjacency does play some part in the explanation of these comprehension phenomena, it remains an open question as to whether the influence of linearity-based (as opposed to hierarchical) relations reflects general properties of the learning and/or processing system, as well as why such linearity effects have not been observed in the comprehension of anaphora.

3.4. How Does the Network Solve the Task?

Our evaluation of the network's behavior still leaves open the question of the nature of the network's generalization concerning anaphora. Has the network discovered abstract constraints like those in (9) and (14), along with the hierarchical relation of c-command? That is, has the network acquired a simple rule-like generalization about the interpretation of reflexives and pronouns, or has it acquired a piecemeal set of facts, with distinct generalizations governing the behavior of different pronouns and reflexives in different structural contexts?

To begin to approach these questions, we conducted an analysis of the patterns of activation of the 70 hidden units in the interpretive network. We reasoned that if the network has derived a uniform generalization concerning the interpretation of reflexives, the activation levels of the hidden units in its interpretive network should be identical for grammatically equivalent sentence contexts. By *grammatically equivalent contexts* we mean positions after which the possible continuations are identical, both from the point of view of word prediction as well as anaphoric interpretation. For example, starting at the point of the * in each of the sentences in (21), the grammar we used to generate our training data will produce identical distribution of words and anaphoric interpretations. In particular, an immediately following reflexive must be interpreted as *John* for each, whereas an immediately following pronoun must not be interpreted as *John*.

(21) a. Simple Matrix: John admires * himself.
b. Object Relative: John who Bill sees admires * himself.
c. Subject Relative: John who sees Bill admires * himself.

Consequently, if the network is treating these contexts identically, as it should, given that the differences regarding elements that intervene linearly between the antecedent and the reflexive are syntactically irrelevant to the anaphoric relation, we might expect that the activation at these points should be identical for these three sentences. To test this hypothesis, we constructed a set of sentences containing examples of all of the three syntactic types illustrated above: simple matrix, subject relative, and object relative. For each of these sentence types, we considered all three

possible names (within a single gender) as the subject and three possible choices for the main verb. For the sentences containing relative clauses, we systematically varied the names within the relative among those of a single gender, but kept the verb within the relative constant. This yielded a corpus of 63 sentences: 9 simple (3 subject names x 3 main verbs), 27 subject relative (3 subject names x 3 main verbs x 3 embedded object names), and 27 object relative (3 subject names x 3 main verbs x 3 embedded subject names).

We then compared, across each of these sentences, the context unit activation patterns that occurred immediately after processing of the main verb. Initial inspection of the context units showed that, contrary to what one might expect, there were large differences in the activation patterns across sentences that differ only in the presence of, or contents of, a relative clause. For one of our networks chosen at random, we found that, within the subset of sentences of the form "Nate ⟨optional relative clause⟩ visits-M himself,'' 31% of the context units had a standard deviation of activation of at least 0.2, 21% at least 0.3, and 3% at least 0.4. (Note that the maximum possible standard deviation for units in the [0,1] range is 0.5.) These differences in activation patterns held not only between sentences that the network processed correctly versus those on which the network made errors, but between the different correct sentences as well (21%, 6%, and 1% respectively).

The large differences in activation patterns raised the question of whether there was any underlying structure among the patterns, reflecting a uniform representation of conditions on anaphora. To begin to approach this question, we applied Hierarchical Cluster Analysis (HCA) to these vectors, using Euclidean distance with complete linkage as the cluster similarity measure. HCA can be used to provide information about the similarity structure of the representational space that the network has developed during learning and has been used to study the representations developed in the context layers of SRNs (Elman 1990, 1995, 1998a). Given a set of points to analyze, HCA builds a tree of clusters of points, recursively merging points into the nearest clusters. The result for one of our networks is shown in Figure 4.

Since only the subject of the sentence is relevant for prediction and anaphoric reference after the main verb, we might expect the patterns to cluster by subject alone. However, the HCA reveals instead that rather than subject, the most important factor determining the distance between patterns was sentence type. First the simple sentences separate from the rest, and the remaining sentences break into two clusters corresponding to subject and object relatives. Relatives are in their own cluster. Within the object relative cluster, the next major division is by subject. Within the subject relative cluster there is no simple second-level division. The fact that the different sentence types are assigned very different activation patterns suggests that the network may be treating the sentence types differently in some way. Of course, the different sentence types must be treated differently while the words comprising the relative clause are being processed. However, after the clause has ended, there is no need to handle the three sentence types separately (and, some would argue, much reason not to).

Not all of the trained networks showed such a clean partition by sentence types; however, such a tendency to group by sentence type was present throughout. We can quantify this tendency by calculating the similarity between the three maximally distant clusters that result from the HCA and the partition of the sentences by subject and by sentence type, using the the Fowlkes and Mallows index (other measures for cluster similarity were highly correlated). The average Fowlkes Mallow index value for similarity to the partition by sentence type was 0.82 (where

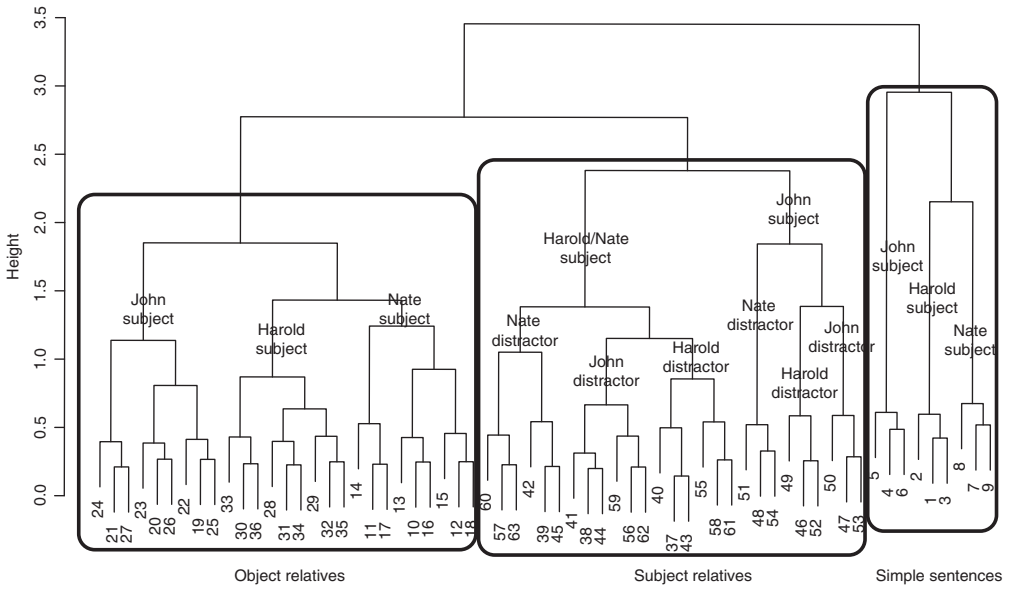


FIGURE 4 Hierarchical Cluster Analysis of network activation immediately preceding the reflexive.

1 represents perfect match), while the average value for similarity to the partition by name was .38, a difference that is significant by a Wilcoxon rank sum test ($p < .01$).

One drawback of HCA is that it will sometimes place points that are actually close together into widely separated clusters. This occurs when the center of a growing cluster "migrates'' away from a data point, leaving it available to be merged into another cluster. This can present a distorted picture of the spatial layout of the points in activation space. Two methods that better preserve distance information are Multidimensional Scaling (MDS) and Principal Components Analysis (PCA). MDS transforms a set of data points into a space of small dimensionality (e.g., 2), in such a way as to preserve the relative distances between the points as much as possible. PCA can be used to find a small set of orthogonal axes that explain a large amount of the variance of the set of data points. One may then examine the layout of points in the space by plotting the points along pairs of these axes. These methods have also been applied to the analysis of connectionist networks (Botvinick & Plaut 2004; Elman 1991; Elman 1993; Elman 1995; Rohde 2002).

We applied MDS to the set of points representing the 63 sentence contexts described above, mapping them into a two-dimensional space. (PCA yielded similar results.) The MDS plot shown in Figure 5 reveals a feature of the representational space that was not evident in the HCA result: The 63 sentences cluster rather into 21 groups, each group corresponding to a conjunction of features; 3 are conjunctions of subject and sentence type (for simple sentences), and 18 are three-way conjunctions of subject, sentence type and name in the relative clause (for subject and object relative).

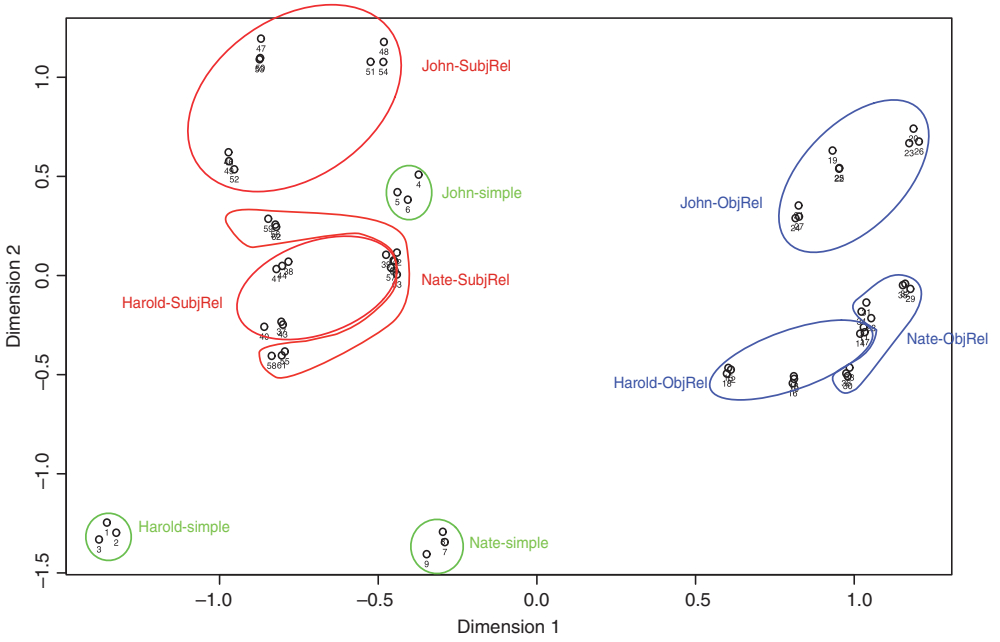


FIGURE 5 Multidimensional Scaling results of network activation of different sentence types immediately prior to reflexive (color figure available online).

This perhaps begins to shed some light on the mechanisms behind the errors the network makes on subject relative sentences: It has not learned to collapse the representation of these sentences across variation in the irrelevant name in the relative clause. Once again, analysis of other networks showed similar structure.

Analyses of hidden-unit activation patterns, such as those just described, can reveal representational distinctions between different domain items. However, those methods do not tell us whether a given representational distinction is actually being used by the network. Some distinctions may simply be the result of initial random weight differences or of differences in input patterns. What constitutes a meaningful distinction in hidden-unit space? Sensitivity (or "lesion'') analysis addresses this question by examining the effects of removing units or connections from the network (Plaut, McClelland & Seidenberg 1995, 1996; Botvinick & Plaut 2004; Allen & Seidenberg 1999).

Recall that our qualitative assessment of the network's performance on sentences including subject relative clauses showed that the network was sensitive to the internal contents of this relative clause along with their arrangement. However, for sentences including object relatives, the network rarely made errors. Does this mean that, when processing this latter class of sentences, the network has (correctly) learned to ignore the contents of the prior relative clause? Or does the network somehow continue to rely on a representation of the relative clause?

To address this question, we compared matched pairs of simple matrix and object relative sentences, which differed only in the presence of a relative clause, e.g., *Nate visits him* vs. *Nate*

who Harold sees visits him. We allowed the network to process each sentence normally, through the main verb. Then, to probe the representation at that point, a single unit in the context layer of the interpretive network was removed, and the network was allowed to process the final word, *him*. We measured the amount by which the error in resolving the pronoun increased as a result of removing the unit. This is a measure of the "importance'' of that unit in the postverb computation for that sentence.

The results were that, for every sentence pair we examined, we were able to find at least one hidden unit whose removal increased the error on the object relative sentence but not on the paired simple matrix sentence. For any such sentence pair, this implies: (i) there is a difference in the representation of the two (equivalent) sentence contexts, because damaging the same component of the representation has different effects on performance on the two sentences; (ii) this difference in representations must be due entirely to the presence of the object relative clause (since that is the only difference between the sentences), and this constitutes a de facto "lingering'' representation of the clause; and (iii) this representational difference is relied upon by the network to perform pronoun reference in the object relative sentence.

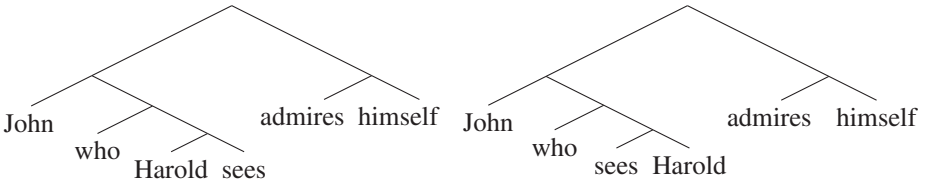
Here we have an example of a network in which lingering representations of grammatically irrelevant clauses are relied upon in resolving subsequent pronouns. This suggests that even in the case of sentences with object relative clauses, where the network is not making errors, the network has not learned the abstract structural dependency between pronouns and their referents that was present in the target grammar.

4. EXPERIMENT 2: STRUCTURAL GENERALIZATION

The fact that the different sentence types are assigned very different activation patterns at a point in the sentence where sentence type is irrelevant is suggestive of the fact that the network's internal structure does not faithfully mirror the generalizations underlying the target grammar. Yet it does not decisively demonstrate that in spite of these representational differences there is not some as yet undetected abstract generalization about anaphoric interpretation that the network is representing in its hidden units.

In order to address this question more directly, we tested the ability of our network architecture to generalize the assignment of reflexive interpretation across the following pair of sentence types.

(22)



Independent of issues of anaphoric reference, the network has other reasons for treating these structures identically at the point following the relative clause: subject-verb agreement and predicting the appropriate distinctive object that goes with the subject. And indeed from the perspective of anaphora, the distinction between these types of relative clauses ought to be irrelevant.

We began this experiment with the intact word-prediction network that was trained in Experiment 1 with sentences of all types. We assume that this network has knowledge of all the sentence types in the original training corpus, that is, simple sentences as well as sentences with each of the two types of relative clauses. We then twice retrained the Phase 2 interpretive network using corpora that were derived from the one used in Experiment 1. The training corpus for No-SubjRel-Net included all of the original sentences but systematically withheld the interpretation of reflexives where the antecedent-reflexive relation spanned a subject relative clause. Similarly, No-ObjRel-Net was trained with the same corpus, but withheld the interpretations of reflexives when the antecedent-reflexive relation spanned an object relative. It is important to emphasize that we did not eliminate subject relative or object relative sentences during Phase 2 training of either of these networks, but simply gave no feedback about the appropriate output for that occurrence of reflexives in that context.

As in Experiment 1, training was carried out using BPTT and no momentum. Batches of 10 sentences were chosen for training by sampling uniformly from the fixed training set, with "online'' updates made to the weights using a cross-entropy cost function. Training was stopped when error on the test sample stopped decreasing. We trained six networks using this regimen, each with different random initial rates in the range $[-1,+1]$.

If in its training the network learns a generalization about reflexive interpretation that cuts across different sentence types, we should expect to see good performance of the network in reflexives in the withheld sentence type. Indeed, this is what Lewis & Elman (2001) found in their study of subject-auxiliary inversion, where the network apparently generalized its knowledge of inversion to withheld constructions. In contrast, if the network learns distinct context-specific generalizations that separately apply to reflexive interpretation in the different sentence contexts, we should expect to see poor performance in the withheld sentence type.

4.1. Results

Over the six simulations, the No-SubjRel-Net achieved mean accuracy of $89.7 \pm 1.5\%$ in reflexive interpretation on a stochastically generated test set of 20,000 sentences that did not include subject relatives. In contrast, mean accuracy on a test set of sentences, all of which included subject relatives, was $60.5 \pm 5.1\%$. The No-ObjRel-Net exhibited a similar disparity in performance: 90.3% accuracy on simple and subject relative sentences and 57.0% on object relative sentences.

4.2. Discussion

The results of this experiment suggest that the network does not acquire a representation of the conditions on reflexive interpretation that cuts across different sentence types but instead learns distinct context-specific generalizations. This conclusion is consistent with what we observed in the network analyses conducted on the network that was given interpretive feedback for the full class of structures.

However, one might object to this interpretation of these results in a number of ways. First of all, one might argue that the testing of the network was unfairly biased against the sentence types for which training data were withheld: The sentences on which the No-SubjRel-Net achieved better performance were on average less complex than those on which it fared well, since the

former set included a significant number of simple matrix sentences, while the latter included only sentences with at least a single relative clause. This observation about the test data is absolutely correct. However, the difference in performance holds up even when we focus on test sets of comparable complexity. For a test set that included all of the sentences generated by the grammar involving a single relative clause, the No-SubjRel-Net achieves 84.1% accuracy on reflexive interpretation for object relatives in comparison to 60.5% accuracy for subject relatives. The No-ObjRel-Net performs in a complementary manner. Thus, the contrasts in performance accuracy between the sampled and withheld sentence types cannot be explained by complexity differences.

An alternative line of objection to our interpretation might accept the claim that the network is indeed learning distinct representations for the different sentence types but only because the network has been overtrained, leading to overfitting. Precisely this sort of overfitting is observed by Elman (2002) with respect to the unlearning of generalizations concerning selectional restrictions. In Elman's experiment, the network begins to learn detailed properties of the training corpus, such as the absence of a particular noun as a possible object of a certain verb, over time leading to a lack of generalization. To test whether such overfitting is occurring in our case, we plotted the performance of the No-SubjRel-Net on reflexive interpretation for the different sentence types over the course of training. This is shown in Figure 6.

As can be readily observed, the network's performance on the withheld example type is considerably below that on the other sentence types throughout training. Thus, there is no basis for concluding that the network has initially learned, but later abandoned, a generalization for reflexive interpretation that cuts across all of the sentence types.

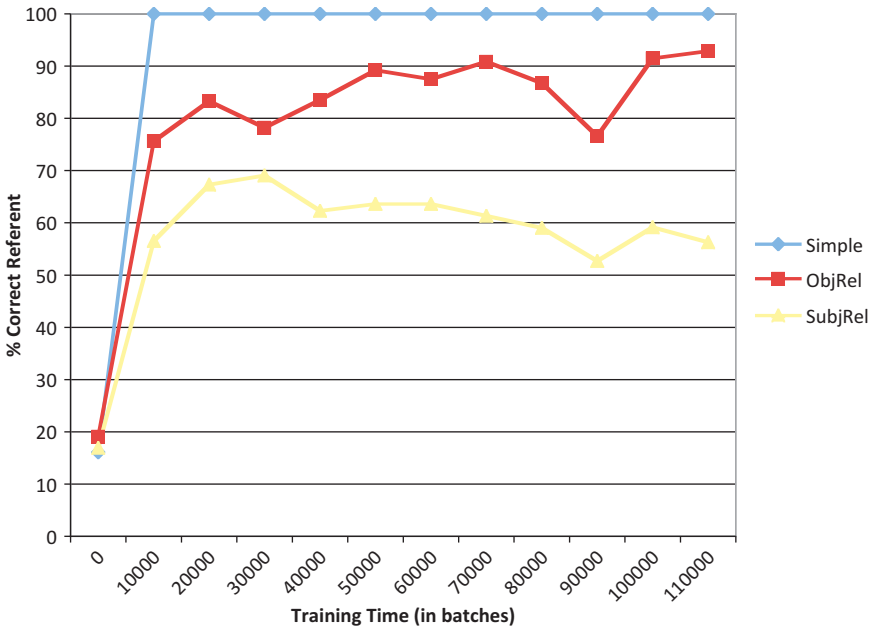


FIGURE 6 Performance of No-SubjRel-Net on reflexive interpretation across training (color figure available online).

5. EXPERIMENT 3: ESTABLISHING ANAPHORIC REFERENCE DIRECTLY

The network architecture from Experiment 1 imposed a severe limitation on the information that can be exploited in the computation of anaphoric reference, namely that which the Phase 1 network represents in order to solve word prediction. As we said above, we pursued this architecture because of the central role that the word-prediction task has played in SRN-based models of language learning. However, one might object that there is no inherent linkage between word prediction and grammar induction per se, and as a result the limitation we have imposed is too draconian. Indeed one might argue that the linearity effects we observed in Experiment 1 are simply an artifact of the kind of architecture and training regimen we employed. In out third experiment, then, we consider a network architecture in which training on the word-prediction and reference tasks is carried out simultaneously and compare its performance to the two-phase architecture from Experiment 1.

The network architecture utilized in this experiment was identical to the Phase 1 network from Experiment 1 with a single difference: An additional set of output units was added, each one providing a localist encoding of the reference of the individuals in the domain. Between this set of output units and the context layer was a hidden layer of 10 units, allowing for the recoding of a distributed representation of anaphoric reference into the localist outputs. This network is depicted in Figure 7.

The training data that we used were identical to those from Experiment 1. Once again six networks were trained using BPTT with a learning rate of .001 and batches of 51 sentences.

5.1. Results

Network performance on the word-prediction tasks was comparable to that seen in Experiment 1. Mean divergence error per word was .011. The average error rate on agreement was 0.1%. Distinctive object prediction, using the same measure discussed above, was accurate at mean rate of 96.7%.

Turning to accuracy on reflexive and pronominal interpretation, we find improved performance. Over our six networks, mean anaphora accuracy on a held-out set of 20,000 (complex) sentences was 98.2%. Accuracy on pronoun interpretation for sentences with a single level of relative clause embedding in modifying the subject was $91.9 \pm 2.4\%$, and $81.5 \pm 2.3\%$ for sentences with two levels of relative embedding.

The contrast between these results and those of Experiment 1 suggests a negative answer to the question raised above concerning the sufficiency of representations deriving from word prediction for computing anaphoric interpretations. Although the domain of sentences on which the networks were trained was constructed so as to require attention to the distinctions necessary for the anaphora task, we see that the representations resulting from training on word prediction are not as well adapted to the anaphora task as representations that derive from explicit and immediate training on the combined task of word prediction and anaphoric reference.

5.2. Linearity Revisited

Given the improved quantitative performance of the network architecture, we can ask whether this is associated with a change in its qualitative behavior. One might want to know, for instance,

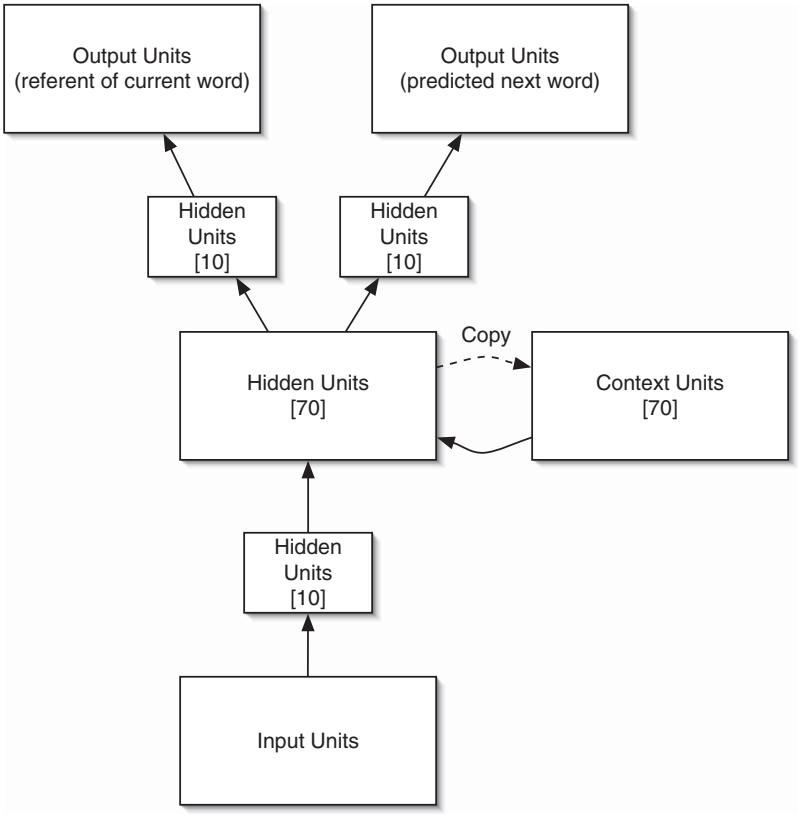


FIGURE 7 Single SRN architecture for anaphoric reference.

whether SRNs trained in a single phase show the same kind of linearity effects as the two-phase networks considered earlier. This question turns out not to be so easy to answer, due to the overall higher level of performance. The spot checks we carried out comparing performance on subject and object relatives does not yield the same sharp contrasts we found in (17) for two-phase networks. Nonetheless, we do find examples with what appear to be detectable linearity effects, like the following:

(23) a. John who loves Mary sees him.
him: $p(Harold) = .52$, $p(Nate) = .48$
b. John who loves Harold sees him.
him: $p(Harold) = .42$, $p(Nate) = .58$

In the first example, the network correctly assigns (essentially) equal probability to the two non-subject male referents as pronoun interpretations. However, when *Harold* forms a linear sequence with the verb and pronoun, as in the second example, we see that there is a slight reduction in the activation of *Harold* as a possible referent for the pronoun, plausibly due to the network's probabilistically taking *Harold* to constitute a subject of a clause *Harold sees him*. This change is

TABLE 4
Mean Activation of Linear Distractor as Reflexive Interpretation (Correct Activation = 0)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.009	.035	497.2	<.001
Net B	.017	.042	465.1	<.001
Net C	.017	.084	1063	<.001
Net D	.035	.047	33.34	<.001
Net E	.039	.090	566.2	<.001

TABLE 5
Mean Activation of Subject Referent as Reflexive Interpretation (Correct Activation = 1)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.984	.952	509.8	<.001
Net B	.971	.914	948.9	<.001
Net C	.967	.869	1456	<.001
Net D	.930	.924	3.883	<.05
Net E	.911	.792	1297	<.001

not nearly as radical as was observed above (cf. (17b)), but it is suggestive that there is an effect of some sort.

To determine the extent of such a linearity effect, we once again analyzed network behavior over a larger class of sentences. As described in section 3.3, we looked at the performance of a number of networks on sentences involving reflexive objects, comparing subject and object relative clauses. For five networks[13] trained using the one-phase training regimen, each with different random initial weights, we again find reliable, though numerically smaller, differences between sentences involving object relatives (which don't induce linearity) and subject relatives (which do) for all networks in all cases. Table 4 shows this for the activation of linear distractor, and Table 5 shows this for the activation of the subject.

For the networks trained under the two-phase training regiment, we did not explore linearity effects for sentences involving pronominal objects, since performance of those networks on the relevant sentences was not sufficiently high for us to conclude that the networks had actually learned the task. Networks trained under the one-phase regimen did considerably better on pronominal reference. We therefore conducted an analysis of linearity effects in these networks for sentences involving pronominal objects. We began by looking at sentences in which the names occurring as main clause subject and within the relative clause were of the same gender, which was in turn identical to the gender of the object pronoun. We labeled this class of sentences XXX, with X a variable ranging over genders. As before, we generated all of the sentences of

[13]We report results here for only five of the then six networks that we trained using this regimen. Data from the sixth network, which was not excluded for any reason related to its behavior, have been corrupted and are no longer available for analysis.

TABLE 6
Mean Activation of Subject Referent as Pronoun Interpretation for XXX Sentences (Correct Activation = 0)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.083	.098	124.3	<.001
Net B	.199	.234	418	<.001
Net C	.195	.186	43.54	<.001
Net D	.168	.195	358	<.001
Net E	.146	.165	126.9	<.001

TABLE 7
Mean Activation of Linear Distractor as Pronoun Interpretation for XXX Sentences (Correct Activation = .5)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.459	.379	509.8	<.001
Net B	.398	.290	4154	<.001
Net C	.395	.353	420	<.001
Net D	.393	.323	1407	<.001
Net E	.404	.331	635.8	<.001

this form that are permitted by the grammar up to two levels of embedding, including both subject relatives and object relatives, a set which numbered 9,300. Our reasoning in this case was parallel to that in the case of reflexives: If a name *N* that is linearly adjacent to the main verb is in some sense taken to be the subject of the sentence, it should this time induce a noncoreference effect for a pronominal object, resulting in lower activation for *N*'s referent as a possible interpretation of the pronoun. Similarly, to the degree that an adjacent noun is taken to be the subject, this should detract from the noncoreference effect induced by the actual subject, assuming noncoreference is typically induced by one referent or another. As a result, we should expect that the referent of the true subject will have higher activation as a possible interpretation of the pronoun in subject relative sentences as compared to object relative sentences. We tested both of these predictions on five separate networks with different random initial weights. In all but one of the networks, the difference in activation of the subjects was highly significant in the predicted direction (though for Net C the numerically small difference was significant in the wrong direction). For the activation of the linearly adjacent name, the difference was highly significant in the predicted direction in all cases. The activation means, and *F*-ratios and *p* values are reported in Tables 6 and 7.

As a second test, we considered sentences of slightly different form: those in which the potential linear distractor differed in gender from the actual subject and the pronoun, which we labeled XYX to indicate that the difference in gender of the distractor. In these cases, since the linearly adjacent name differs in gender from the object pronoun, it will not contribute any additional anticoreference effect. However, we should still expect that the activation of the subject's referent as a pronoun interpretation should be greater for subject relative sentences to the degree to which the linear distractor is taken to be a subject in the relevant sense. Here, this difference is significant in four out of five networks, as shown in Table 8.

TABLE 8
Mean Activation of Subject Referent as Pronoun Interpretation for XYX Sentences (Correct Activation = 0)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.080	.122	1393	<.001
Net B	.191	.214	235.4	<.001
Net C	.157	.160	8.749	<.01
Net D	.164	.184	318.7	<.001
Net E	.151	.148	3.706	n.s.

TABLE 9
Mean Activation of Linear Distractor as Pronoun Interpretation for XYY Sentences (Correct Activation = .333)

	ObjRel	*SubjRel*	*F-ratio*	*Prob*
Net A	.321	.283	528	<.001
Net B	.328	.265	1461	<.001
Net C	.334	.306	367	<.001
Net D	.352	.287	2341	<.001
Net E	.340	.276	1392	<.001

As a final test in the pronoun domain, we considered a third class of sentences, labeled XYY, where the linear distractor agrees in gender with the object pronoun but differs from the subject. Here, we should expect no effect on the activation of the subject's referent, which should remain at 0 because of the gender mismatch between subject and pronoun, and indeed no such effect was observed. However, we should see a noncoreference effect on the object pronoun in the face of a linearity effect: The activation of the linear distractor's referent should be lower in subject relative sentences than in object relative sentences. Once again, this effect is strongly supported, this time in all of the networks, as shown in Table 9.

Each of the five networks studied here shows signs of systematic linearity effects for a majority of the cases we looked at. Moreover, of the 30 comparisons we carried out, we found only a single instance of an "antilinearity'' effect, where the linearly adjacent name (in the object of a subject relative clause) is less effective than a nonlinearly adjacent name (in the subject of an object relative clause) in inducing noncoreference effects (Net C in Table 6). This prevalence of linearity effects leads us to conclude that the change from a two-phase training regimen to one involving only a single phase of training has not affected the network's predisposition to attend to linear-based patterns.

5.3. Structural Generalization in Single-Phase Networks

Thus far, we have seen that the single-phase training regimen, when compared to the two-phase training we considered earlier, SRN has substantial quantitative, if not qualitative, effects on SRN performance in assigning pronominal and anaphoric reference: the resulting networks perform at

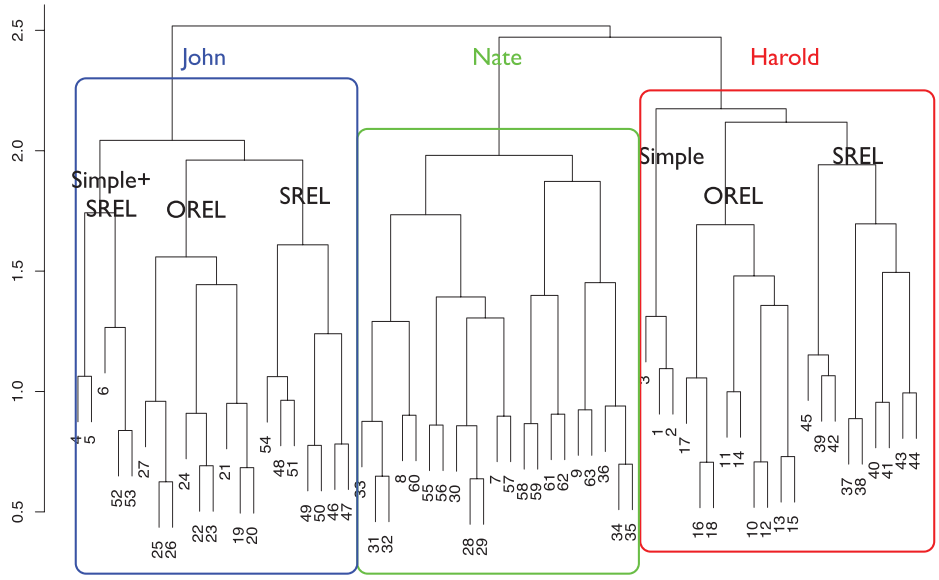


FIGURE 8 Hierarchical Cluster Analysis for one-phase SRN (color figure available online).

a higher level of accuracy but still show linearity effects. Let us now revisit the issue of structural generalization. Recall that when we analyzed the hidden-unit activations of the two-phase network immediately after the processing of the main verb, we found substantial differences according to sentence type (cf. Figure 4). If we perform the same analysis on one-phase networks, the results are quite different. Figure 8 shows the result for one of our networks.

This HCA diagram suggests that the network is representing information about a sentence's subject's referent, a property that impacts subsequent processing for both word prediction and anaphoric interpretation, more robustly than the subject's structural type, a property that does not. Performing HCA on the other networks trained using this regimen yields similar results, even if not always so sharp. As previously, we can quantify the degree to which three maximally distinct clusters are organized according to subject identity or sentence type using the Fowlkes and Mallows index. In this case, the average Fowlkes Mallow index value for similarity to the partition by sentence type was only .43 (where 1 represents perfect match), while the average value for similarity to the partition by name was .75, a difference that is significant by a Wilcoxon rank sum test ($p < .05$). Comparison between the values obtained from one-phase and two-phase networks for sentence type and for subject yields significance in both cases, again by the Wilcoxon rank sum test ($p < .01$ for both sentence type and subject). This suggests that the one-phase networks have a consistent bias to represent information about subject more robustly than information about the sentence type, while two-phase networks show the opposite bias.

These results leave open the question of whether the network can take advantage of this similarity in representation to generalize anaphor interpretation from one sentence type to another. To answer this question, we once again trained the SRN depicted in Figure 7, this time withholding target outputs for the referents of reflexives when the subject was modified by a subject

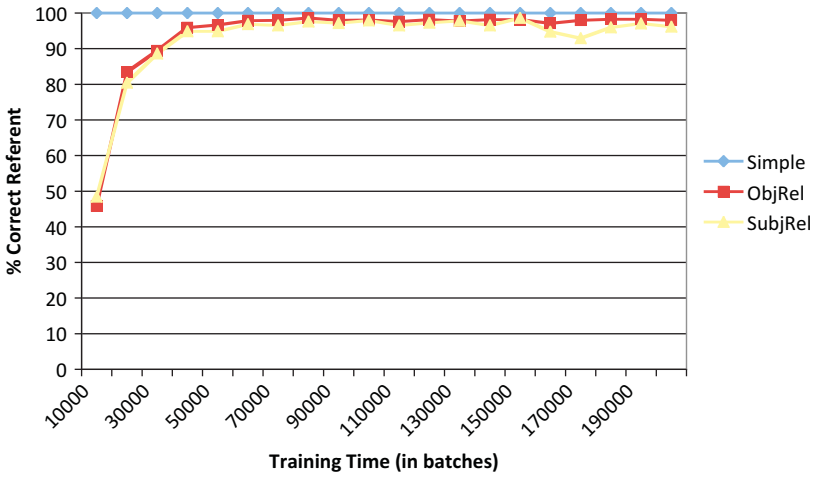


FIGURE 9 Structural generalization in one-phase SRN (color figure available online).

relative clause. As before, training was carried out using BPTT and no momentum. Batches of 10 sentences were chosen for training by sampling uniformly from the fixed training set, with "online'' updates made to the weights using a cross-entropy cost function. Training was stopped when error on the test sample stopped decreasing.

The resulting network generalizes quite differently from the one derived from the two-phase training regimen considered previously. As seen in Figure 9, accuracy on the withheld sentence type increases along with the others as training progressed, though its final performance was somewhat lower.

Recall that the two-phase networks studied in Experiment 2 did not show such generalization (cf. Figure 6). The difference between these two generalization patterns was highly significant. This contrast suggests that there is something about the two-phase training regimen that led to distinct representations for different sentence types.

Given the results of the HCA just discussed, this result is not very surprising: Similarly represented sentences tend to be treated alike, and hence the assignment of an interpretation for reflexives in the context of subjects modified by subject and object relatives will be treated through the same set of connection weights. If we take this diagnosis seriously, it suggests that there may continue to be sharp limits on the structural generalizations that this network can make. Specifically, in the HCA in Figure 8, simple and complex sentences for the same subject referent are still represented quite differently one from the other. As a result, we need not expect robust generalization across these sentence types.[14] To test this prediction, we carried out two additional experiments. In the first, we withheld the target outputs for the referents of all reflexives whose

[14]Of course, there may be similarities between these representations that are being clouded by other distinctions, and these (clouded) similarities could nonetheless provide the network with a basis for generalization across sentence types. Nonetheless, to a first degree of approximation, the kind of representational distinctiveness detected by HCA often diagnoses behaviorally relevant representational distinctions, and we therefore consider whether this is true in this case.

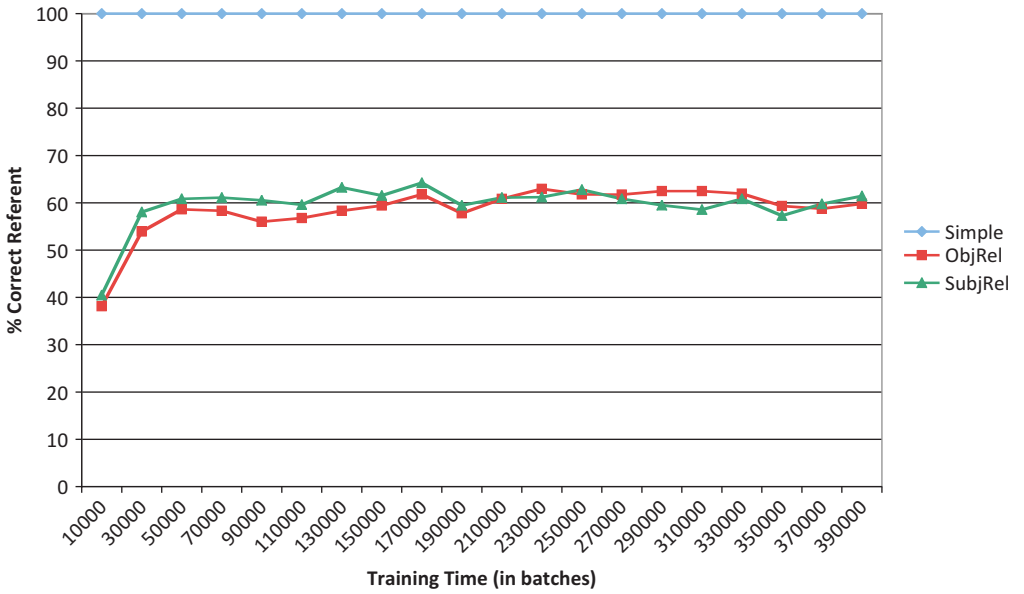


FIGURE 10 Generalization of reflexive interpretation from simple to (held-out) complex sentences in one-phase SRN (color figure available online).

subjects were modified by a relative (either subject or object), while in the second, we withheld the referents of all reflexives with simple subjects. In both cases, the antecedents of reflexives inside of relative clauses (e.g., *John who adores himself sees Lucy*) were retained. In order to succeed at this task, the network needs to generalize its knowledge of simple sentences to more complex ones. When trained with held-out complex subjects, four networks with different random initial weights achieved an average maximum accuracy of 66% at assigning a reference to a reflexive whose antecedent is a complex subject, while achieving perfect performance on simple subjects. This maximum occurred at different points in training for each network, and in computing this average we took the maximum for each network separately. Figure 10 gives the average performance for the three sentence types across these networks during training.

With held-out simple subjects, four trained networks achieved an average maximum accuracy of 66% on reflexives whose antecedents are simple subjects, while achieving nearly perfect performance on the complex subject cases. Again, maximum accuracy occurred at different points for each of the networks. The average performance of these networks during training is shown in Figure 11.

Though imperfect, we see then that these one-phase networks are capable of some degree of generalization across sentence types.

Given the success of these networks, we considered a further modification in attempt to increase generalization still further: a reduction in the number of hidden units. There is a widely noted tendency that networks with fewer hidden units generalize better to novel inputs because of their more limited capacity to represent detailed properties of the training inputs that would distinguish them from the test data. Of course, along with a reduction in the number of hidden units

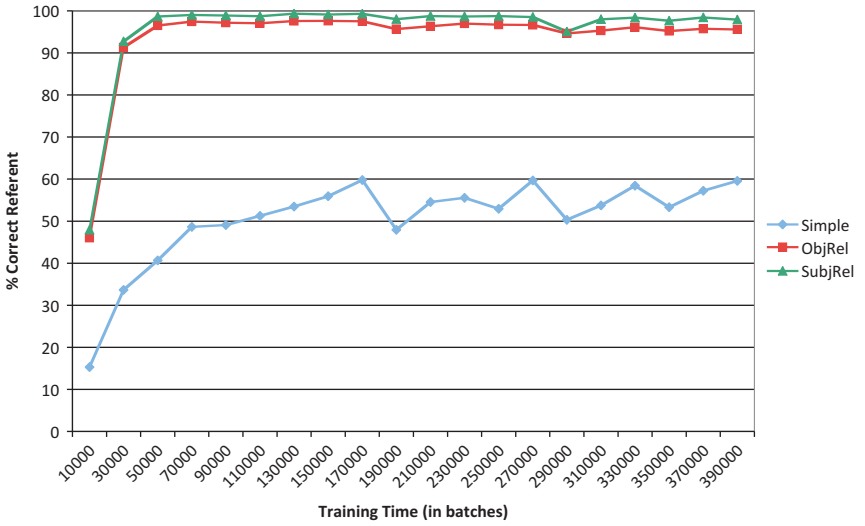


FIGURE 11 Generalization from complex to (held-out) simple sentences in one-phase SRN (color figure available online).

comes the possibility that the network will be less capable of solving the task, even if it treats trained and novel inputs uniformly. We therefore explored whether a reduction in the size of the hidden-unit layer would improve generalization on anaphoric interpretation for novel structures while retaining an overall high level of performance for the anaphoric interpretation and word prediction tasks. We considered networks with a range of sizes: 50 hidden units, 30 hidden units, 15 hidden units, and 7 hidden units. We trained two networks of each size with different random initial weights in each of the generalization conditions studied here. With the exception of the 7 hidden-unit case, all of the networks achieved comparable performance on word prediction, subject-verb agreement, and unique object prediction to our original 70 hidden-unit networks. Since the 7 unit network's performance was reduced overall, we omitted it from further consideration. For the held-out complex sentence condition, the average maximum accuracy on anaphoric interpretation for the held-out sentences was 73% with 50 units, 82% with 30 units, and 67% with 15 units, the first two of these being substantially higher than generalization performance in the original 70 hidden-unit networks, which was 66%. This suggests that 30 units is a nearly optimal network size for learning this task. Indeed, in the held-out simple sentence condition, average maximum accuracy for the held-out sentences was 100% for the 30 hidden-unit networks. Thus, it appears that a reduction in the representational capacity of these SRNs can facilitate generalization of anaphoric interpretation to novel structural contexts, even if generalization from simple to complex sentences does not achieve the ceiling levels seen in generalization from complex to simple or from subject relative sentences to object relative sentences (see Figure 9).

An interesting property of these results is that generalization is more successful from complex to simple sentences rather than the reverse. This is in contrast to what a number of models of syntax learning (Lightfoot 1989; Wexler & Culicover 1980) have suggested, where evidence gleaned from simpler sentences is used to determine the properties of more complex ones. It is intriguing that SRNs seem to show a bias in favor of extending generalizations in the opposite fashion.

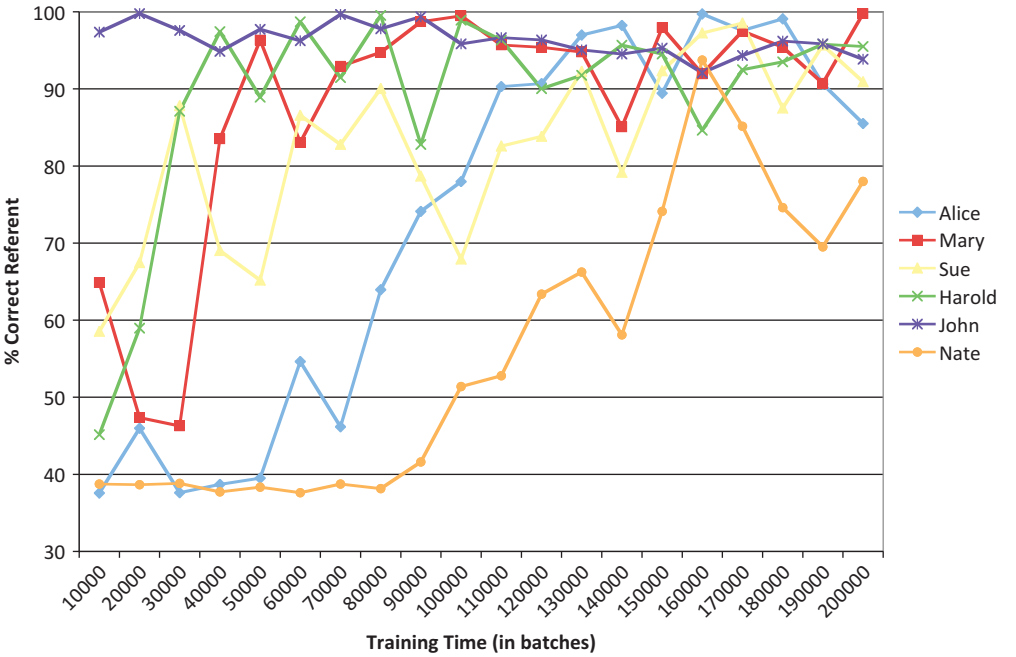


FIGURE 12 Learning of pronoun interpretations by subject name for one-phase SRN (color figure available online).

6. EXPERIMENT 4: LEXICAL GENERALIZATION

One question left unanswered by our previous experiments concerns another kind of abstraction in linguistic knowledge, namely that which cuts across different lexical items. This arises with the regularities governing reflexive and pronoun interpretation, which are independent of the specific antecedent involved, but depend only on the context in which the pronoun or reflexive is found. To what degree does the network's knowledge of the anaphoric interpretation cut across different names? One first bit of evidence bearing on this question derives from the learning curve of the network considered in Experiment 3. In Figure 12, we see the time course of learning for pronoun interpretation. Each curve in the figure represents the network's accuracy over the course of training in assigning the correct referential possibilities to a pronoun for a set of sentences in which the target antecedent is one of the six names in the domain (possibly modified by a relative clause).

The figure strikingly displays the degree to which the network acquires its knowledge of the referential possibilities for pronouns in a piecemeal fashion: Particularly for the last three names in the domain, the network seems to hold off on learning the conditions on (non)coreference for pronouns that are c-commanded by that name in the same clause until the interpretive possibilities for the previous name are fully mastered. Of course, this learning curve by itself does not prove that there is not a single unified generalization concerning pronominal interpretation at

the conclusion of training: it is possible that as training progresses the interpretive conditions on pronouns become increasingly general.

To test the degree to which this is true, we explored an SRN's capacity to induce an interpretive mapping for reflexive interpretation that generalizes across names, a sure indication that anaphoric reference across names is treated by the network in a uniform fashion. We utilized the SRN architecture studied in Experiment 3, using the single-phase training regimen that achieved best performance and structural generalization. We trained five networks using different random initial weights. As in the previous experiment, we used the training corpus from Experiment 1, but this time withheld interpretations for reflexives on the basis of reference. Specifically, we withheld training on the reference for all reflexives whose interpretation was *John*. Note that it is not the case that sentences with *John* as antecedent of a reflexive never occurred during the training of the network. Rather, the networks were simply not given feedback during training about the appropriateness of its interpretive output for reflexives in sentences of this sort.

6.1. Results

For nonwithheld names, the networks achieved performance comparable to that in the previous studies. On a stochastically generated test set, the networks on average assigned the correct referent to the reflexive in 99.4% of the cases. For the withheld name, however, the networks' performance was strikingly different. Four of the five networks never correctly assigned the correct referent for the reflexive, even in simple sentences such as *John saw himself*, while one assigned it correctly 2.2% of the time. This pattern held throughout training, as shown in Figure 13, which depicts the test set performance during training.[15]

6.2. Diagnosing the Failure

The complete failure to generalize to the held-out name recalls Marcus's (1998) "a rose is a rose'' problem, albeit in a more cognitively natural context.[16] Marcus demonstrated that an SRN trained to do word prediction on a corpus of sentences like *a lily is a lily* and *a tulip is a tulip* (all of the form *an X is an X*) is unable to generalize the pattern to a novel nouns such as *rose*. That is, after training, a network that had not been trained on the sentence *a rose is a rose* would not correctly predict the next word, namely *rose*, after receiving as input the sequence *a rose is a*. As Marcus points out, this follows directly from the nature of backpropogation learning: Since the connections to the relevant input and units are modified independently one from the other during training, the absence of experience with a novel word will mean that the connections to and from that word will never be modified beyond the random initial settings. As a result, the network cannot generalize to the novel word any acquired knowledge that is instantiated in its connections. Marcus uses the term *training space* to refer to those inputs that are within the cross-product of the feature values (represented by the individual input and output units) that are within the training data and argues that feedforward networks and SRNs are incapable of generalizing

[15] Attempts to facilitate generalization by reducing the number of hidden units, as discussed in the last section, had no effect on the outcome in this case.

[16] Indeed Marcus (1998:266) already noted the similarity of his "a rose is a rose" problem to that of establishing a relation between a reflexive and its antecedent.

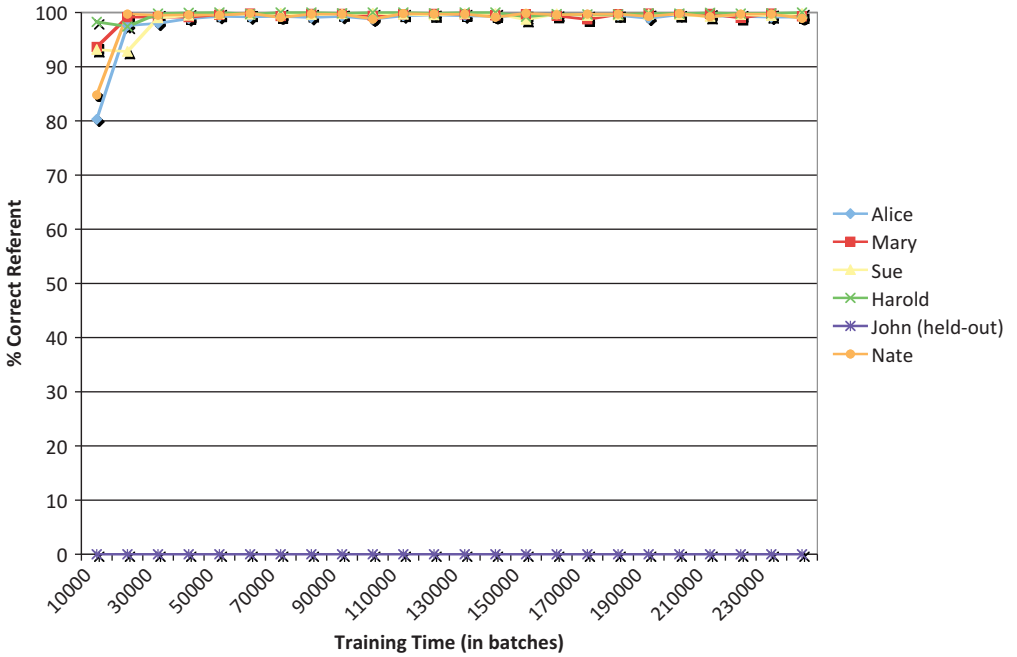


FIGURE 13 Test set performance on reflexive interpretation by name during training (color figure available online).

beyond the training space for the reasons of training independence just discussed. Since the word *rose* is not included in the training set, any sequence containing the word *rose* will fall outside of the training space, and as a result the network will not be able to systematically generalize to such a sequence.

Marcus's demonstration might strike the reader as limited in scope. If a network has no experience whatsoever with the noun *rose* whose localist input representation for *rose* shares nothing with other words on which the network was trained, it is unfair to expect that the network could have knowledge of the contexts in which *rose* should be predicted. Note, however, that our experiments with anaphora are not subject to this sort of training space argument. During training, the network is presented with the reflexive *himself* on the input side, as well as with the reference unit corresponding to *John* (as the reference for the name *John*) on the output side.[17] Moreover,

[17]Marcus discusses a modification of his "a rose is a rose" experiment that also sidesteps the training space issue, but where the network nonetheless fails. To the previous training set, he adds sentences containing the previously withheld word *rose* that are not of the form *an X is an X* (the example he gives in the article is *the bee sniffs the rose*). In the face of such examples, the network does have reason to modify its connections from and to the input and output units corresponding to the word *rose*. However, even with this modification, the SRN still fails on the "a rose is a rose" task. Marcus suggests that this result nonetheless follows from his training space conception. He proposes that prediction in the different sentence types constitutes distinct functions that the network must compute, and the notion of training space is defined "with respect to some particular function." Though Marcus's intuition about multiple functions is reasonably clear, it is not at all obvious what it would mean for the network to divide its task into separate functions or, even in the

the "held-out'' sentences are not completely held out: only the target activations for the reference units are withheld, so that no weights are updated at that point during the processing of the sentence. We might imagine that the predication task could lead the network to develop distributed representations that encode commonalities across names, and this commonality could be exploited in defining the conditions on anaphoric reference. Nonetheless, as we have seen, the network fails to generalize.

An anonymous reviewer suggests that the failure in generalization might arise not from the absence of experience with particular units in the input and output but from combinations of these units. That is, even though *himself* occurs as an input word and *John* occurs (independently) as a referent output, the training data never require that they be associated one with the other. If such an explicit association were required by SRNs in order to succeed (something we do not believe to be the case in human learning), it would point to a rather severe limitation on their ability to generalize. Nonetheless, we tested whether adding such an association to the training data is sufficient to change the behavior of the network. In order to do this, we needed to somehow add sentences in which *John* is the referent of *himself*, but which are not instances of sentences having *John* as the subject and *himself* as the object, as these are the very sentences to which we want to see generalization. One way to do this would be to introduce a novel name, say *Jack*, into the domain, whose referent was identical to that of *John*, and include sentences of the form *Jack admires himself* in the training set. When we do this, what happens is that already at the first layer of hidden units the network collapses its representation of the words *John* and *Jack*, so that they are indistinguishable at the recurrent layer and beyond. Unsurprisingly, then, the network successfully generalizes to sentences with *John* as the antecedent of *himself*, but for uninteresting reasons. To avoid this problem, we added to our training set sentences that included pronominal subjects. Each occurrence of a pronominal subject was assigned an interpretation that was any of the gender appropriate referents in the domain, including the withheld referent *John*. Since the single pronoun *he* could take on a number of distinct referents, the network could not collapse its representation with that of any of the names. In accordance with the usual constraints, a reflexive object in a sentence with a pronominal subject was assigned an interpretation that was identical to the subject of the sentence. Thus, in the sentence *he admires himself*, if the interpretation of the pronoun was assigned to be *Nate*, so too was the interpretation of the reflexive. And since *John* was a possible interpretation for *he* in the training data, it also occurred as a possible interpretation for the reflexive *himself* as well. This modification of the training data had no effect on the outcome of our experiment. The network failed completely to identify *John* as the antecedent for the reflexive in sentences like *John sees himself*. Thus, even the addition of explicit associations between input and output units is not sufficient to induce generalization.

In fact, we believe any characterization of the limits of SRN generalization cast purely in terms of the input and output units alone is doomed to fail. Unlike feedforward networks, the output of an SRN depends not only on the input units that are activated at a given point in time, but also on the activation of the hidden units at the previous time step. A network may fail if its representation of the previous context is inadequate in some respect, even if each of the words in the context has itself occurred. However, requiring that the network be exposed to all possible

face of a coherent definition of such a notion, why we should expect that the network would divide it along the lines that seem most natural to the experimenter. Thus, without considerable analysis of the network's functioning, it will be difficult indeed to advance a function-specific training space argument.

contexts for a particular word would be tantamount to giving up on generalization. The crucial question, it seems to us, is how the network chooses, during learning, to represent the context, and whether this representation collapses contexts that are relevantly similar and provides access to the appropriate degrees of freedom.

In the face of the profound failure of generalization we find here, we still might wonder about the locus of the failure. Has the network completely neglected to generalize its knowledge of anaphoric dependencies to a novel name, or is its failure of a more local nature? To get at this question, we analyzed the network's hidden-unit representations at the point in processing in which the reflexive is presented to the network. In particular, we considered the pattern of activation in the network's penultimate 10 unit hidden layer that is connected to the interpretive output units. If the network is to succeed in the task of assigning the correct reference to a reflexive, the representations in this layer must appropriately distinguish among the sentential contexts according to the name that occurs as the subject (since this determines the appropriate antecedent). As Minsky & Papert (1969) show, these representations must be linearly separable: Unless one can find a (10-dimensional) hyperplane to separate the regions of (10-dimensional) activation space that correspond to each of the referential possibilities, the single layer of connections to the output units will not permit the network to distinguish among them appropriately. In order to get at this question, we constructed a set of sentences, all with reflexive objects, but with different subjects, with and without both subject and object relative clauses, and with different verbs. We ran two of the networks from Experiment 4 on these sentences and extracted the activation vectors at the penultimate hidden layer. We then classified these vectors according to the name occurring as the subject of the sentence in the preceding context. To determine linear separability, we trained a simple perceptron to reproduce the classification of vectors by antecedents. The result was that the perceptron was successful in learning this task, indicating that these hidden unit representations were indeed linearly separable.

It is tempting to conclude from this result that the network has indeed learned something about reflexive interpretation that has generalized to the new name, but for some reason yet to be understood has not been able to exploit the representational distinction in the mapping to the output units. Before we can do this, however, we must be sure that the linear separability of the activation vectors according to antecedent is the result of learning by the network and not simply the by-product of a rich representational space where a vast number of partitionings of the activation vectors are linearly separable.

In order to test this latter hypothesis, we did two things. First, we investigated whether the networks' activation vectors in the penultimate hidden layer after processing the reflexive were linearly separable when partitioned according to a variety of properties of the sentential context (in addition to the antecedent of the reflexive): whether the sentence had a simple subject, whether its subject was modified by an object relative clause, whether the subject was one of the male nouns, whether the previous word in the sentence was a specific verb (*admires*), whether the first word in the sentences was a particular name (*Nate*), and whether the name occurring most recently (in linear order) is a particular name (again *Nate*). In addition, we studied the degree to which the activation vectors that arise in the network prior to training, given its random initial weights, are linearly separable when partitioned by the same properties.

The results, which are given in Table 10, show two things. First, the substantial number of negative results points to the fact that the representational space of the hidden unit vectors at this point in the network architecture is not so rich as to permit arbitrary properties of the sentential

TABLE 10
Linear Separability of Penultimate Hidden Layer Activation Vectors at *Himself*

	Reflexive Antecedent	*Simple Sentence?*	*ObjRel Sentence?*	*Male Subject?*	*Previous Word is* admires*?*	*First Word in Sentence is* Nate*?*	*Most Recent Name is* Nate*?*
Pretraining	✘	✔	✘	✔/✘	✘	✘	✘
Posttraining	✔	✘	✘	✔	✘	✘	✘

TABLE 11
Linear Separability of Context Layer Activation Vectors at *Himself*

	Reflexive Antecedent	*Simple Sentence?*	*ObjRel Sentence?*	*Male Subject?*	*Previous Word is* admires*?*	*First Word in Sentence is* Nate*?*	*Most Recent Name is* Nate*?*
Pretraining	✘	✔	✘	✔	✘	✘	✘
Posttraining	✔	✔	✔	✔	✘	✔	✔

context to be distinguished. Interestingly, if we consider the linear separability of the same set of properties in the network's context layer, the results are different, as seen in Table 11.

A second and more intriguing conclusion that we can draw from the results in Table 10 is that linear separability by reflexive antecedent is a property of the network's representations that is acquired during training. This means that during training the network's weights have been adjusted so that the activation vectors associated with contexts in which the reflexive *himself* should be interpreted as *John* are distinctive, in spite of the fact that this pairing was held out during training. From this, we can infer that some aspect of the network's knowledge about anaphoric reference has generalized to a name that was novel in this context. The question that remains for the future is why the network was unable to exploit the representational distinction that it has acquired in the final mapping to the output units and what modifications to the training regimen or network architecture might allow it to do so.

7. CONCLUSION

To sum up, we see that when judged in quantitative terms SRNs are remarkably successful in learning to map pronouns and reflexives onto their grammatically possible antecedents. However, we have found that the manner in which this success is achieved diverges from the abstract structural conditions that have been proposed in the linguistics literature. The results of Experiments 1 and 3 point to a bias on the part of SRNs in favor of linear-based generalizations even when such generalizations do not supply categorically reliable cues to interpretation in the training data. Moreover, we also saw in Experiments 2, 3, and 4 that the networks did not systematically induce lexically or structurally abstract generalizations for anaphoric reference, though we did see signs of structural abstraction under the single-phase training regimen. Instead, these results pointed to

the conclusion that SRNs, given the sort of data we presented in these experiments, have a tendency to induce construction-specific and lexically specific generalizations rather than lexically neutral generalizations in terms of abstract hierarchical relations such as c-command. These networks resist extending the interpretive mapping for reflexives and pronouns beyond those specific structures and referents on which they had been trained.

This limitation appears to be in serious conflict with human performance: For example, we know who *himself* refers to in *Gromit sees himself* even without prior experience with a reflexive having Gromit as referent or even without having any prior knowledge of who the name *Gromit* refers to. (See Marcus 2001 and Hadley 1994 for arguments that humans exploit this kind of ability to universally generalize about individual entities in a variety of domains.) Likewise it would seem problematic that for the networks studied here, there would appear to be no evidence that the complementary interpretations for the pronoun and reflexive in *John admires him/himself* reflects the biases of SRN learning: if it had been otherwise, one might have anticipated that if John were the withheld reflexive interpretation, it would nonetheless benefit from that fact that it had succeeded in learning that John could not be the interpretation assigned to the corresponding pronoun.

Moving to a rule-based model of anaphoric interpretation, whether restricted by hard-wired constraints (Pearl & Lidz 2009) or learned entirely from data (Soon, Ng & Lim 2001), would of course allow for the induction of lexically and structurally abstract generalizations and could be formulated in such a way as to resist linearity-based generalization via an appropriate prior. And while we suspect that such an approach may well form the basis of the solution to the learning of anaphoric dependencies, we remain curious as to whether there are non-rule-based approaches that can achieve lexical and structural abstraction.

One potentially fertile area for exploration involves the use of distributed rather than localist representations. As has been pointed out in the connectionist liternature, the use of distributed representations permits a network to generalize across related elements, by exploiting commonalities in their featural makeup. Indeed Elman (1998b) hypothesizes that the use of distributed output representations would help SRNs overcome some of their limitations. However, there are significant technical problems to making this work, some noted by Marcus (1999). Since the output of a word-prediction SRN represents a probability distribution over possible next words, such a distribution would need to be represented via the simultaneous activation of all of the component features for each of multiple lexical outputs, leading to a "superposition catastrophe,'' an inability to extract the distinct lexical items in the output and their relative probabilities from the multiple features that are active. Alternatively, a static output pattern could be used to represent an SRN's predictive distribution over words, but this would allow only a very restricted class of distributions to be represented. We believe that this hurdle can be overcome via a novel architecture in which the output layer of an SRN is replaced by a Restricted Boltzmann Machine (RBM, Smolensky 1986). Activation of an RBM fluctuates over time in a way that can be used to represent a probability distribution over distributed representations via the temporal likelihood of the RBM being in a certain representational state. Such a hybrid SRN-RBM architecture could simultaneously achieve distributed representations and arbitrary output distributions by using a stochastic RBM to realize the output distribution. RBMs can be easily trained to assign highly dissimilar probabilities to highly similar patterns but can nonetheless take advantage of similarity in pattern similarity in the absence of explicit evidence to the contrary. In the case of lexical generalization we have considered here, although lexical items may be held out of the training

set, using distributed representations would ensure that no network output units are held out of the training set, and as a result the network could learn the conditions under which the output features should and should not be activated. This learning about individual features can carry over from one lexical item to the next, since each is composed of multiple features. The SRN-RBM architecture directly allows the learning of relationships between features of input words and features of output words that is required for just such generalization. Pilot simulations have shown that networks mapping features-to-features in this way can learn, for example, an identity mapping between distributed input and output representations, which allows the network to correctly activate lexical items that it has never seen before in the training corpus. Whether this would constitute a solution that is generalizable to the full range of cases remains to be seen.

Finally, we point out that there have been a number of proposals in the acquisition literature that suggest that learning what come to be abstract grammatical generalizations starts off being tied to specific lexical items and constructions (Goldberg 1998; Tomasello 1992; Tomasello 2000). Therefore the results of the experiments reported here showing a lack of lexical and structural generalization might be taken as demonstrating that SRNs provide an appropriate inductive mechanism for this first stage of grammatical development. What remains to be determined is whether one can move beyond this stage within the context of distributed network computation in the form of an SRN by modifying the training regimen or architecture in some fashion, perhaps building on the tantalizing results of the analysis in Experiment 4, or whether progress demands a different approach to mental computation involving the symbolic representation of grammatical knowledge, or whether the best solution is one that merges ideas from network and symbolic computation (see, for example, Chang 2002).

ACKNOWLEDGMENTS

For much useful discussion, suggestions, and criticisms, we are grateful to Esteban Buz, Janet Fodor, Christo Kirov, Jeff Lidz, Jeff Markowitz, Lisa Pearl, William Sakas, Paul Smolensky, John Stowe, our anonymous reviewers, as well as audiences at numerous colloquium and conference presentations. Financial support was provided by National Science Foundation (NSF) grant SBR-0446929. The material in this article is based in part on work done while the third author was employed as an NSF Program Director. Any opinion, findings, and conclusions expressed in this article are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation.

REFERENCES

Allen, Joseph & Mark S. Seidenberg. 1999. The emergence of language: The emergence of grammaticality in connectionist networks. In Brian MacWhinney (ed.), *The emergence of language*, 115–152. Hillsdale, NJ: Lawrence Erlbaum Associates.

Asudeh, Ash & Frank Keller. 2001. Experimental evidence for a predication-based binding theory. In Mary Andronis, Christopher Ball, Heidi Elston & Sylvain Neuvel (eds.), *Papers from the 37th annual meeting of the Chicago Linguistic Society*, vol. 1, 1–14. Chicago, IL: Chicago Linguistic Society.

Badecker, William & Katherine Straub. 2002. The processing role of sructural constraints on the interpretation of pronouns and anaphora. *Journal of Experimental Psychology: Learning, Memory and Cognition* 28. 748–769.

Botvinick, Matthew & David C. Plaut. 2004. Doing without schema hierarchies: A recurrent connections approach to normal and impaired routine sequential action. *Psychological Review* 111. 395–429.

Chang, Franklin. 2002. Symbolically speaking: A connectionist model of sentence production. *Cognitive Science* 26. 609–651.

Chomsky, Noam. 1986. *Knowledge of language, its nature, origin, and use*. New York: Praeger.

Christianson, Kiel, Andrew Hollingworth, John F. Halliwell & Fernanda Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology* 42(4). 368–407.

Cleeremans, Axel, Bert Timmermans & Antoine Pasquali. 2007. Consciousness and metarepresentation: A computational sketch. *Neural Networks* 20(9). 1032–1039.

Crain, Stephen & Mineharu Nakayama. 1987. Structure dependence in grammar formation. *Language* 63(3). 522–543.

Duffy, Susan A., John M. Henderson & Robin K. Morris. 1989. Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory and Cognition* 15(5). 791–801.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14. 179–211.

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

Elman, Jeffrey L. 1993. Learning and development in neural networks: The importance of starting small. *Cognition* 48. 71–99.

Elman, Jeffrey L. 1995. Language as a dynamical system. In Robert F. Port & Tim van Gelder (eds.), *Mind as motion: Explorations in the dynamics of cognitiion*, 195–236. Cambridge, MA: MIT Press.

Elman, Jeffrey L. 1998a. Generalization, simple recurrent networks, and the emergence of structure. Paper presented at the Symposium on Congnitive Architecture at the twentieth annual conference of the Cognitive Science Society, University of Wisconsin, Madison.

Elman, Jeffrey L. 1998b. Representational issues: Commentary on *The Algebraic Mind*. http://www.psych.nyu.edu/gary/TAM/elman.html (24 May, 2013).

Elman, Jeffrey L. 2002. Generalization from sparse input. In Mary Andronis, Erin Debenport, Anne Pycha & Keiko Yoshimura (eds.), *Proceedings of the 38th annual meeting of the Chicago Linguistics Society*, vol. 2, 175–200. Chicago, IL: Chicago Linguistic Society.

Elman, Jeffrey L., Elizabeth Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi & Kim Plunkett. 1996. *Rethinking innateness*. Cambridge, MA: MIT Press.

Ferreira, Fernanda, Kiel Christianson & Andrew Hollingworth. 2001. Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research* 30. 3–20.

Fischer, Silke. 2004. Optimal binding. *Natural Language and Linguistic Theory* 22. 481–526.

Fisher, Cynthia. 2002. Structural limits on verb mapping: The role of abstract structure in 2.5-year-olds' interpretations of novel verbs. *Developmental Science* 5(1). 55–64.

Fodor, Jerry A. & Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28. 3–71.

Francis, W. Nelson. 1964. *A standard sample of present-day English for use with digital computers*. Report to the U.S. Office of Education on Cooperative Research No. E-007. Providence, RI: Brown University.

Godfrey, John J., Edward C. Holliman & Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 1, 517–520. Washington, DC: IEEE Computer Society.

Goldberg, Adele E. 1998. Patterns of experience in patterns of language. In Michael Tomasello (ed.), *The new psychology of language*, 203–219. Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, Peter C. & Randall Hendrick. 1997. Intuitive knowledge of linguistic co-reference. *Cognition* 62. 325–370.

Hadley, Robert F. 1994. Systematicity in connectionist language learning. *Mind and Language* 9. 247–272.

Hadley, Robert F. 2003. Systematicity of generalization in connectionist networks. In Michael A. Arbib (ed.), *Handbook of brain theory and neural networks*, 1151–1156. Cambridge, MA: MIT Press.

Haskell, Todd R. & Maryellen C. MacDonald. 2005. Constituent structure and linear order in language production: Evidence from subject–verb agreement. *Journal of Experimental Psychology: Learning, Memory and Cognition* 31(5). 891–904.

Hemforth, Barbara & Lars Konieczny. 2003. Proximity in agreement errors. In *Proceedings of the 25th annual conference of the Cognitive Science Society*, 557–562. Mahwah, NJ: Lawrence Erlbaum Associates.

Joanisse, Mark & Mark S. Seidenberg. 2003. Phonology and syntax in specific language impairment: Evidence from a connectionist model. *Brain and Language* 86. 40–56.

Konieczny, Lars. 2005. The psychological reality of local coherences in sentence processing. In *Proceedings of the 27th annual conference of the Cognitive Science Society*. csjarchive.cogsci.rpi.edu/proceedings/2005/docs/p1178.pdf.

Koster, Jan & Eric Reuland (eds.). 1991. *Long-distance anaphora*. Cambridge: Cambridge University Press.

Lewis, John D. & 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*, 359–370. Somerville, MA: Cascadilla Press.

Lightfoot, David. 1989. The child's trigger experience: Degree-0 learnability. *Behavioral and Brain Sciences* 12. 321–334.

MacWhinney, Brian & Catherine E. Snow. 1985. The child language data exchange system. *Journal of Child Language* 12. 271–296.

Marcus, Gary F. 1998. Rethinking eliminative connectionism. *Cognitive Psychology* 37. 243–282.

Marcus, Gary F. 1999. Response to Jeff Elman's commentary on *The Algebraic Mind*. http://www.psych.nyu.edu/gary/TAM/author_response.html (24 May, 2013).

Marcus, Gary F. 2001. *The algebraic mind*. Cambridge, MA: MIT Press.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Macinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19. 313–330.

McClelland, James L. & Karalyn Patterson. 2002. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Science* 6. 465–473.

Minsky, Marvin & Seymour Papert. 1969. *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press. (Expanded edition 1988).

Nicol, Janet, Kenneth Forster & Csaba Veres. 1997. Subject-verb agreement processes in comprehension. *Journal of Memory and Language* 36. 569–587.

Nicol, Janet & David Swinney. 1989. The role of structure in co-reference assignment during sentence processing. *Journal of Psycholinguistic Research* 18. 5–19.

Pearl, Lisa. 2011. When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. *Language Acquisition* 18(2). 87–120.

Pearl, Lisa & Jeffrey Lidz. 2009. When domain general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development* 5(4) 235–265.

Pearlmutter, Neal J., Susan M. Garnsey & J. Katherine Bock. 1999. Agreement processes in sentence comprehension. *Journal of Memory and Language* 41. 427–456.

Pinker, Steven & Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28. 73–193.

Pinker, Steven & Michael Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Science* 6. 456–463.

Plaut, David C., James L. McClelland & Mark S. Seidenberg. 1995. Reading exception words and pseudowords: Are two routes really necessary? In Joseph P. Levy, Dimitrios Bairaktaris, John A. Bullinaria & Paul Cairns (eds.), *Connectionist models of memory and language*, 145–159. London: UCL Press.

Plaut, David C., James L. McClelland, Mark S. Seidenberg & Karalyn Patterson. 1996. Understanding normal and impaired word reading: Computation principles in quasi-regular domains. *Psychological Review* 103. 56–115.

Pollard, Carl & Ivan Sag. 1992. Anaphors in English and the scope of binding theory. *Linguistic Inquiry* 23(2). 261–303.

Reinhart, Tanya & Eric Reuland. 1993. Reflexivity. *Linguistic Inquiry* 24. 657–720.

Rodriguez, Paul. 2001. Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation* 13(9). 2093–2118.

Rodriguez, Paul, Janet Wiles & Jeffrey L. Elman. 1999. A recurrent neural network that learns to count. *Connection Science* 11. 5–40.

Rohde, Douglas. 1999a. A connectionist model of sentence comprehension and production. Pittsburgh, PA: Carnegie Mellon University Ph.D. thesis.

Rohde, Douglas. 1999b. The Simple Language Generator: Encoding complex languages with simple grammars. Tech. Rep. CMU-CS-99-123. Pittsburgh, PA: Carnegie Mellon University, Department of Computer Science.

Rohde, Douglas. 2002. A connectionist model of sentence comprehension and production. Pittsburgh, PA: Carnegie Mellon University dissertation.

Rohde, Douglas & David C. Plaut. 1999. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition* 72. 67–109.

Rumelhart, David E. & James L. McClelland. 1986. On learning the past tenses of English verbs. In James L. McClelland & David E. Rumelhart (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 2, chap. 18. Cambridge, MA: MIT Press.

Runner, Jeffrey T., Rachel S. Sussman & Michael K. Tanenhaus. 2003. Assignment of reference to reflexives and pronouns in picture noun phrases: Evidence from eye movements. *Cognition* 89. B1–B13.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory. In David E. Rumelhart, James L. McClelland & the PDP Research Group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1: *Foundations*, 194–281. Cambridge, MA: MIT Press.

Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* 27(4). 521–544.

Sprouse, Jon & Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's *core syntax*. *Journal of Linguistics* 48(3). 609–652.

Sturt, Patrick. 2003. The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language* 48. 542–562.

Tabor, Whitney, Bruno Galantucci & Daniel Richardson. 2004. Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language* 50. 355–370.

Tomasello, Michael. 1992. *First verbs: A case study of early grammatical development*. Cambridge: Cambridge University Press.

Tomasello, Michael. 2000. Do young children have adult syntactic competence. *Cognition* 74(3). 209–253.

Wexler, Kenneth & Peter W. Culicover. 1980. *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

Wilson, Colin. 2001. Bidirectional optimization and the theory of anaphora. In Géraldine Legendre, Jane Grimshaw & Sten Vikner (eds.), *Optimality-theoretic syntax*, 465–507. Cambridge, MA: MIT Press.

Xiang, Ming, Brian Dillon & Colin Phillips. 2009. Illusory licensing effects across dependency types: ERP evidence. *Brain & Language* 108. 40–55.

Yang, Charles D. 2004. Universal grammar, statistics or both? *Trends in cognitive sciences* 8(10). 451–456.

Submitted 31 January 2012
Final version accepted 28 August 2012