

Running head: ACQUISITION OF ANAPHORA BY SRNS

# The Acquisition of Anaphora by Simple Recurrent Networks

Robert Frank, Donald Mathis, and William Badecker

Johns Hopkins University

**Abstract**

This paper describes experiments in applying the Simple Recurrent Network (SRN) architecture (Elman 1991, 1993) to the task of learning the grammatical dependencies involved in reflexive and pronominal anaphora. This task requires more refined sensitivity to grammatical structure than those that have been explored in previous explorations conducted with SRNs. When evaluated quantitatively, SRNs performed very well in assigning reference to pronouns and reflexives. However, we found that the manner in which SRNs achieved such performance and the generalizations that the networks learned diverged in certain key respects from those of the target grammar, leading to systematic and errors unlike those made by humans.

# 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 by which the word order in a language is made to depart from the canonical sequence of phrasal groupings through the preposing or postposing 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 the adherence to structural principles governing possible coreference relations between phrases; 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, many 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 and/or universal the property can be shown to be.

Over the past 20 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 to extract generalizations. As Elman et al. (1996) have emphasized, the connectionist approach does not deny the importance of innately provided inductive bias, but 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.

This line of inquiry has obvious appeal, and part of this appeal is that the kinds of innate structure necessary under the connectionist perspective seem more biologically plausible than that 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 how it is that precisely 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 forming the English past tense. 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 entirely upon the levels of activation at the input units (along with the network connectivity), SRNs include recurrent connections which 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 sentence. 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. Rodriguez et al. (1999) have shown that an SRN trained in this fashion is capable of inducing the kinds of structural 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 transitive 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 simple 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, as there is no fixed distance or 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 a hierarchically structured representation of the sentence, 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 successful results. Lewis and Elman (2001), for instance, show that SRNs can learn the conditions on subject-verb inversion, inducing the correct structure-based rather than linear-order based generalization (cf. Crain and Nakayama (1987)). Rohde (1999a) demonstrates that the SRNs can acquire knowledge of the conditions under which *want to* can contract to *wanna*, conditions which 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 paper lays out our first explorations of both of these issues, by focusing on the ability of SRNs to extract the generalizations concerning the interpretation of anaphoric elements, specifically pronouns and reflexives.<sup>1</sup> 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 will allow 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 learners do not acquire the conditions under which a reflexive like *herself* may refer to an individual Mary separately from the conditions under which it may refer to an individual Sue. Instead, they appear to learn a set of conditions that constrain interpretation for occurrences of the reflexive with any possible antecedent.<sup>2</sup> Reflexive interpretation, then, can

---

<sup>1</sup>Joanisse and 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.

<sup>2</sup>This is different from what has been claimed, by Tomasello (1992, 2000) among others, in the context of the learning of subcategorization frames and argument structure alternations. Here it has been argued that such structures are learned on a verb by verb basis. See Fisher (2002) for a different view. Whatever the resolution of this issue, we take it to be uncontroversial that the acquisition of anaphora does not work in this way, and indeed it is conspicuous

be taken to involve the use of a rule containing a variable, which may be instantiated by any structurally accessible antecedent. By structurally abstract, we mean that the language learners acquire generalizations about the possible interpretations for a reflexive element that cut across (irrelevant) structural factors. For instance, the possibility of interpreting a reflexive in the direct object of a verb as coreferent with the subject does not 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 questions:

- 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?
- 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 variable-containing generalizations, an issue that has been hotly debated by Marcus (2001) and Elman (1998b).

## 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*

---

that to our knowledge such a proposal has never been made in this domain.

and *himself* show restrictions on the noun phrases from which they take their interpretation, their *antecedents*. For instance, the reflexive direct object *herself* in (1) may take as its antecedent *Mary's mother*, but not *Mary*.

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

Similarly, the reflexive *himself* in (2) may have as its antecedent only the noun phrase *the man who knows John*, and not *John*.

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

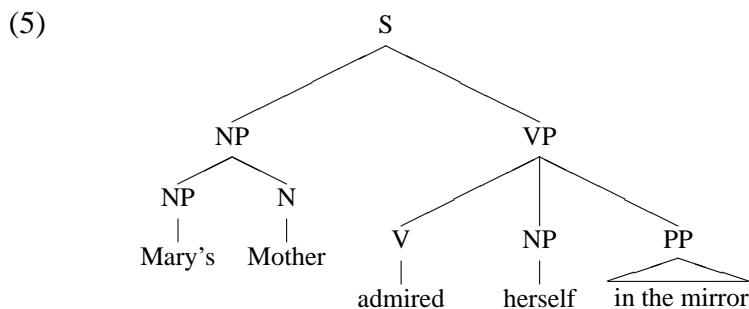
Such restrictions have been characterized in terms of the structural relation called *c-command*, which can be defined for present purposes as follows:

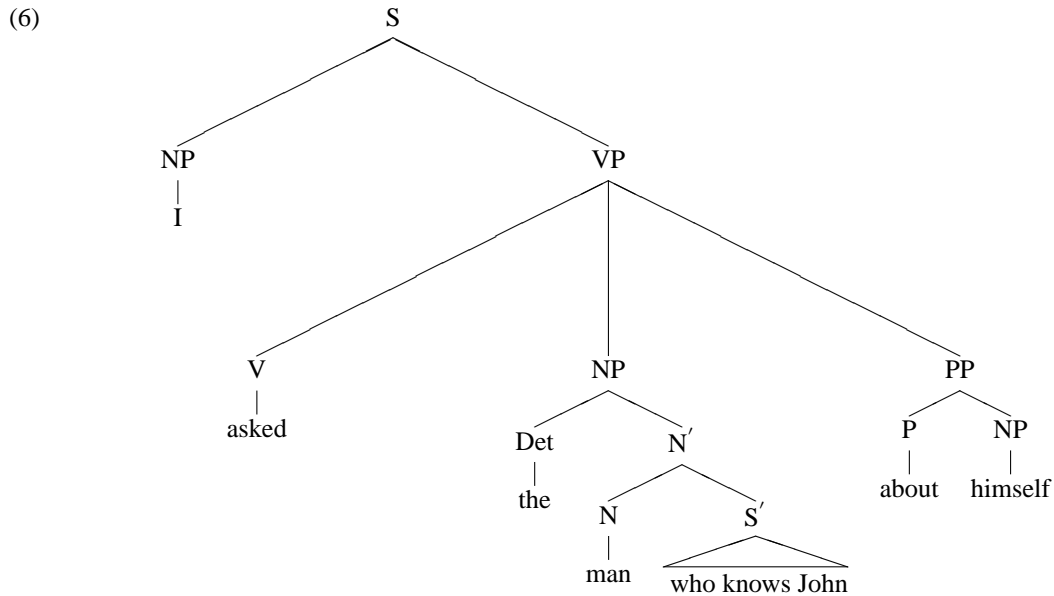
- (3) X c-commands Y iff the mother of X dominates Y and X does not dominate Y.

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

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

If we look at the standardly assumed structural descriptions for the sentences in (1) and (2), as shown in (5) and (6) respectively, we see that the possible antecedents for the reflexives all stand in a relation of c-command: in the first case, the mother of the NP *Mary's mother* is the S, which dominates the reflexive, and in the second case, the mother of the NP *the man who knows John* is VP, which dominates the reflexive.





In contrast, none of the impossible antecedents c-commands the reflexive, as the reader can readily confirm. The c-command condition is not sufficient to characterize the possible antecedents for a reflexive. For instance, in (7), *herself* can refer only to Alice in spite of the fact that the NP *Sue* c-commands it as well.

(7) 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:

(8) 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:

(9) Mary’s mother admired her in the mirror.

(10) 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 (9) can refer to any female individual except Mary’s

mother (even someone not mentioned in the sentence), and the pronoun *him* in (10) can refer to any male individual except the man who knows John. This pattern points to the following constraint on English pronoun interpretation.

(11) 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.

(12) 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 above 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 reflexive's clause. We can therefore revise our condition on pronoun interpretation as follows:

(13) 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 the notion of locality in a more precise way for the condition on either pronoun or reflexive interpretation, and the experiments we describe will not probe this aspect of anaphoric interpretation.<sup>3</sup>

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 (8) and (13) are similar those one finds in linguistics texts and are usually taken to represent part of the abstract system of linguistics knowledge a speaker possesses. Such a model of linguistic competence is

---

<sup>3</sup>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 and Reuland 1991). We put aside discussion of this issue for the remainder of this paper, 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.

to be distinguished from a model of linguistic performance, which would specify the mechanisms that underlie on-line processing. Since connectionist modeling eschews the difference between competence and performance, or, stated less contentiously, aims at directly modeling human processing, one might argue that these abstract constraints on pronoun and reflexive interpretation are simply irrelevant. Instead, on this view one should be studying and directly characterizing patterns of human sentence processing. Whether or not we accept this rejection of the notion of linguistic competence and the importance of modeling it, it is important to note that with respect to the anaphoric phenomena under discussion, they simply do not diverge. Experimental results from Gordon and Hendrick (1997), Asudeh and Keller (2001), Badecker and Straub (2002) and Sturt (2003) uniformly confirm the sensitivity to a c-command-based condition in pronoun and reflexive interpretation, along the lines of those in (8) and (13).<sup>4</sup> Of course, abstract structural constraints like these 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 with the output of on-line sentence processing, it strikes us as a reasonable first step to try to model the patterns of interpretation they generate.

---

<sup>4</sup>It is true that these experimental results as well as those from Runner et al. (2003) do find limited contexts in which a divergence between on-line processing and abstract grammatical principles is observed. Note however that all of the reported cases of divergence are related to the locality portion of the constraints on anaphoric interpretation rather than the c-command based condition, and therefore are not directly relevant to the issues explored in this paper. Additionally, the cases in which divergence occurs, so-called “picture NPs” are notorious even within the theoretical literature, and are known to pose difficulties for a range of theoretical proposals.

- (i) Joe saw Ken’s picture of himself.

Thus, it would be overstating the empirical situation rather substantially to say that these experimental results falsify the proposals made in the context of studies of competence grammar.

### 3 Experiment 1: Establishing anaphoric reference

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 and specifically of subject-verb agreement was assessed in terms of its success at predicting the verb 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 agreement. Knowledge of anaphoric interpretation, in contrast, cannot be completely assessed via word prediction. In this task, the knowledge of hierarchical grammatical structure that the network acquires during training must be put to use not only in predicting the next word (though it is certainly relevant for prediction, for example to know 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 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 about the task of making predictions that allows the network to induce a representation of hierarchical structure for the sentence. If word prediction depends on identifying the actual structural relations that exist between the next word and the words that precede it, then the hidden unit representations of such a network should suffice for learning the structurally defined conditions on anaphoric interpretation. Following this logic, 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-

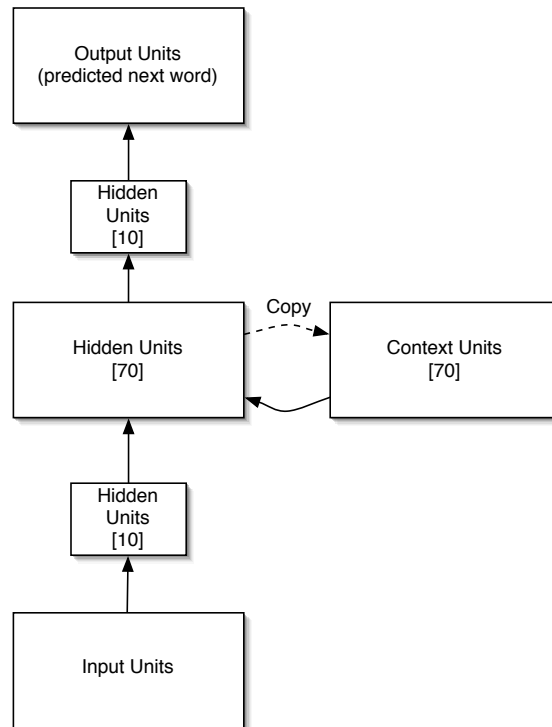


Figure 1: Phase 1 network

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. and
- a recurrent “context” layer of 70 units was placed between the pair of hidden layers. At each time step in processing, the activation levels of these units were copied to a set of 70 addition units. At the next time step, these activation levels were provided as input to the original 70 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 interpret the activation values at the outputs as the networks predicted probability distribution for the next word.

We assume, following previous work, that this first training phase on the word prediction task will suffice to allow the network to acquire a hierarchical representation of the sentences in the training corpus. If this representation is of the sort that linguists have found necessary for characterizing grammatical phenomena such as constraints on anaphoric interpretation, it ought to suffice as input to a network that assigns interpretation. Following this reasoning, the weights of phase one network were frozen prior to the initiation of the second training phase. Then, the hidden units of the original phase one network were used as input 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 reference to each word in the sentence, using as input the hidden units of the phase one network. We assumed a locality representation of interpretation. Specifically, each word in our language other than reflexives and pronouns was associated with a unique output unit that was designated as the uniquely active interpretive output unit when the corresponding word was presented as input. 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 (8) always determined a unique possible antecedent for the reflexive. For pronouns, this was not the case. The constraint in (13) tells us what may not function as the antecedent of the pronoun, but does not say what must. Thus in general the choice

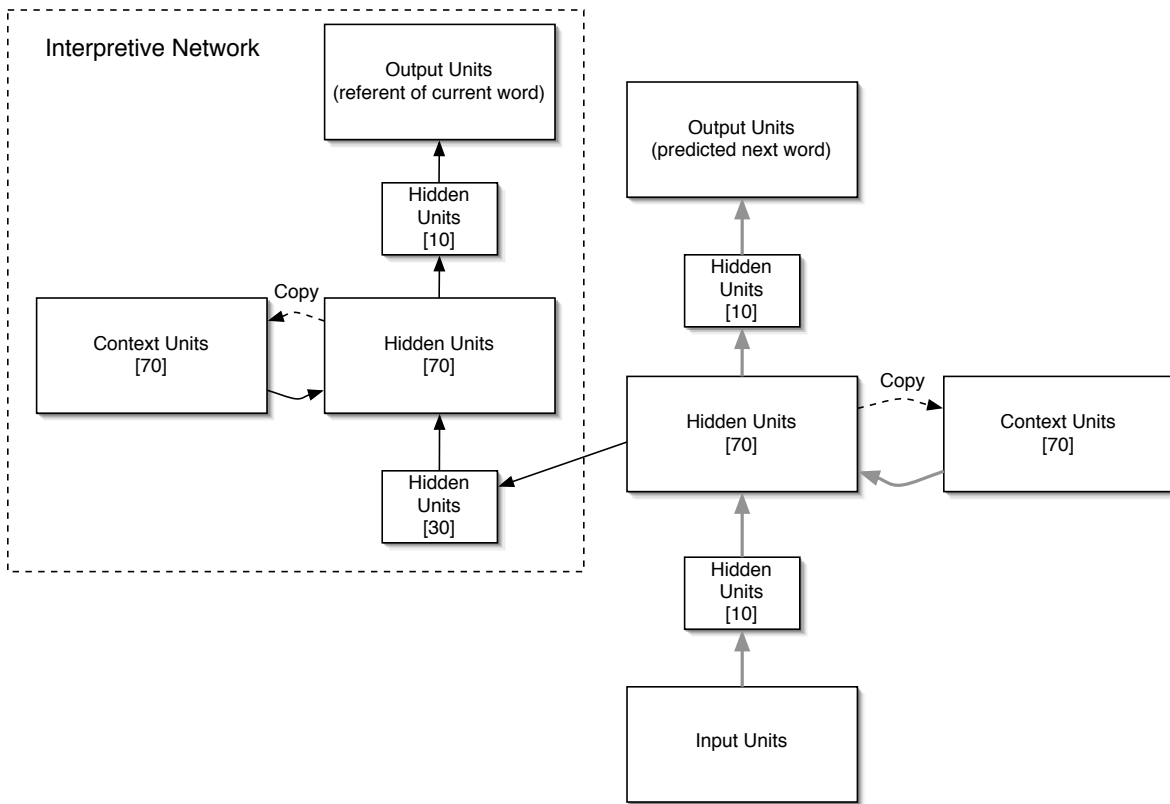


Figure 2: Phase 2 network

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. For each occurrence of a pronoun in some sentence of our training data, a particular referent was randomly selected among those that were compatible with the restriction in (13), and the network was trained with this referent as the target for that sentence. It is worth keeping in mind this non-determinism of the pronominal reference task, as we might expect the network to have more difficulty.

### 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. 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 possible antecedent. The language, then, included sentences of the following forms:

(14) 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

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

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

*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 can be thought of as more naturalistic in the sense that it incorporated admittedly coarse-grained semantic restrictions on word co-occurrence (cf. Rohde and Plaut (1999)). This addition to our language also has the effect of requiring the network to represent the identity of a name in order to carry out word prediction in the subsequent structural context, as it will determine possible unique object noun phrases. Otherwise, the word prediction network had 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.

The language described in the previous paragraph was described using the grammar shown in figure 3. Using the Simple Language Generator tool described in Rohde (1999b), a number of 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 relatives with subject and object gaps. Among direct objects, the probability of a reflexive was 12.5%, as was the probability of a direct object pronoun. We then stochastically generated a training corpus of 18,000 sentences and a test set of 2,500 sentences from this language. These sentences had a

minimum length of 3 words and were truncated to a maximum length of 21 words (average length was 6.2 words). Of these, approximately half (58%) 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 next word predictions for simple sentences. However, Rohde and 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 our simulations, we follow Rohde and Plaut in taking the initial weight range to be  $[-1, +1]$ , thereby eliminating the need to stage the training.

Both phase one and phase two networks were trained using backpropagation through time (BPTT) and no momentum.<sup>5</sup> 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, which occurred after 160,000 updates.

## 3.2 Results

We quantitatively evaluated the performance of the network in a number of ways.

---

<sup>5</sup>Strictly speaking, the SRN model, as originally proposed by Elman, makes use of a different learning algorithm from the one we are using, namely standard backpropagation. We are however interested in understanding what the range of generalizations is that the SRN architecture can detect. Since BPTT is a more powerful learning algorithm than standard backprop, we adopt BPTT in order to give the SRN architecture the best chance of learning the target grammars. Further, we follow Rodriguez in continuing to call this an SRN architecture.

**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 and Plaut (1999), we assessed the difference between these two distributions in terms of Kullback-Leibler divergence, a measure of the relative entropy of two distributions. 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 was .029. This is very slightly worse than the results reported by Rohde and Plaut (1999). However, it should be noted that the addition of the distinctive objects makes the task of word prediction considerably more difficult.

**Agreement accuracy** To ensure that our phase one 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 of verbal agreement. An error was scored if a single output unit representing a non-agreeing 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. The error rate on agreement was 0.1% on novel sentences.

**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, the network achieved 97.3% accuracy. Quantitative assessment of the accuracy on pronominal reference is more difficult because there is no deterministic solution to the problem, but our examination of a wide range of examples in a stochastically generated training corpus turned up very few errors. This suggests on first blush that the network has been quite successful at solving the reference task, and that the hidden unit representation derived through word prediction is indeed structurally rich enough to support interpretive processes.

### 3.3 Behavioral analysis

These quantitatively measures suggest at first blush that the network has been quite successful at solving the reference task, and that the hidden unit representation derived through word prediction during phase one training is rich enough to support interpretive processes, and provides the necessary inductive bias to lead the network to identify the appropriate generalization. However, as impressive as these quantitative successes are, it is also important to note that the network's behavior in the anaphoric interpretation task diverged in certain systematic ways from what one would expect from English speakers. Consider, for instance, the networks performance on the sentences in (15).

- (15) a. Nate who Harold likes admires him

$$him: p(Harold) = .59, p(John) = .38$$

- b. Nate who likes Harold admires him

$$him: p(Harold) = .30, p(John) = .35, p(Nate) = .33$$

In both of these examples, the judgments and on-line processing data of English speakers points to the unavailability of *Nate* as an antecedent for the pronoun *him*, in accordance with the grammatical constraint in (13) since *Nate* c-commands *him*. As indicated by the probabilities listed below

example (15a), the network correctly interprets this case, assigning high probability to both *Harold* and *John* as possible antecedents of *him*. But for (15b), which has the same structural relation between *Nate* and *him*, the network assigns probability mass to *Nate*, incorrectly, since *Nate* is the subject of the sentence of which *him* is the object.

From the perspective of the kind of linguistic analyses of anaphora reviewed in section 2, this contrast is puzzling as 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 (15b) of a linear sequence that forms a possible simple sentence, *Harold admires him*. In this sequence, if it were taken to be an independent sentence, *Nate* would be a possible antecedent for *him* (although *Harold* would not), and we suggest that the network is basing its response on the union of the antecedents that would be possible structurally and linearly. (In other simulations we have run, the network appeared to choose a strategy of intersecting the reference possibilities, giving only *John* as a possible antecedent for sentence like (15b). It should be noted, however, that in all of the networks we have trained there have been exceptions to these patterns). In example (15a), there is no such linear sequence, and as a result the network correctly bases its decision on structural factors. Example (16a) shows that we find similar effects when the intervening name is feminine. Here, the presence of the sequence *Mary admires him* licenses interpreting the pronoun as referring to any of the male individuals in the domain.

(16) a. Nate who sees John who kisses Mary admires him

*him*:  $p(\textit{Harold}) = .28, p(\textit{John}) = .33, p(\textit{Nate}) = .38$

b. Nate who sees John who kisses her admires him

*her*:  $p(\textit{Alice}) = .38, p(\textit{Mary}) = .26, p(\textit{Sue}) = .34$

*him*:  $p(\textit{Harold}) = .53, p(\textit{John}) = .43$

We see the linear effect dissolve in (16b), this time because the name is replaced by a pronoun, thereby forming a sequence, *her admires him*, that is not a possible sentence.

Similar effects of linearity were also observed in reflexive interpretation, but exclusively with sentences having two levels of embedding:

(17) Nate who sees Harold who kisses John admires himself

*himself*:  $p(\text{John}) = .29$ ,  $p(\text{Nate}) = .70$

As before, we see that the network gives as possible interpretations the union of those which should be possible on the basis of structural and linear factors (i.e., the presence of the sequence *John admires himself*).

Note that the presence of these linear effects does not show that the phase one network's hidden units provided an insufficiently rich representation of the syntactic structure to support the task of anaphoric reference. On the contrary, the response patterns just exhibited can be interpreted as demonstrating a sensitivity to syntactic structure: In (17), for instance, we assume that *Nate* is taken to be a possible antecedent of the reflexive (in contrast to *Harold*) precisely because it is the subject of the main clause, and therefore c-commands the reflexive. Indeed, we take the presence of this pattern, in which the subject remains active, to defuse an alternative potential explanation for the difficulty of such cases for the network, deriving from the difficulty of maintaining activation across intervening material. Nonetheless, the presence of these linearity effects does however 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. As far as we are aware, such a pattern is never attested in human processing of anaphora.<sup>6</sup> We hypothesize that this dis-

---

<sup>6</sup>This stands in contrast to the phenomenon of subject-verb agreement, where intervening elements that are not in subject position can nonetheless have effects on the conditioning of agreement:

- (i) The key to the cabinets were missing.

Even here, however, it is not clear that the relevant property of the false “attractor” of agreement is linear adjacency. For recent discussion, see Franck et al. (2002), Haskell and MacDonald (2003) and Hemforth and Konieczny (2003). Assuming that linear adjacency is part of the explanation of this phenomenon, it remains an open question as to why

inction 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.

### 3.4 Network analysis and discussion

Our qualitative evaluation of the network's performance still leaves open the question of the nature of the network's generalization concerning anaphora. Has the network discovered abstract constraints like those in (8) and (13), 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 its knowledge in a piecemeal fashion?

To begin to approach these questions, we conducted an analysis of the patterns activation of the 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 (18) 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, an immediately following pronoun must not be interpreted as *John*.

- (18) 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 context 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

---

such linear effects have not been observed in anaphora.

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 compared, across each of these sentences, the context unit activation patterns that occurred immediately after processing 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 which differ only in the presence of, or contents of, a relative clause. For example, for the subset of sentences of the form “Nate ⟨optional relative clause⟩ visits-M himself”, we found that 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 vs. 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 as the dissimilarity 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 results are shown in Figure 4.

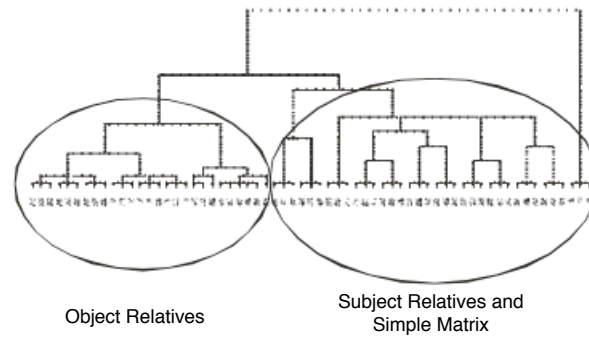


Figure 4: Hierarchical Cluster Analysis of network activation immediately before reflexive

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 analysis revealed that instead of 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.

One drawback of HCA is that it will sometimes place points which 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 which 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

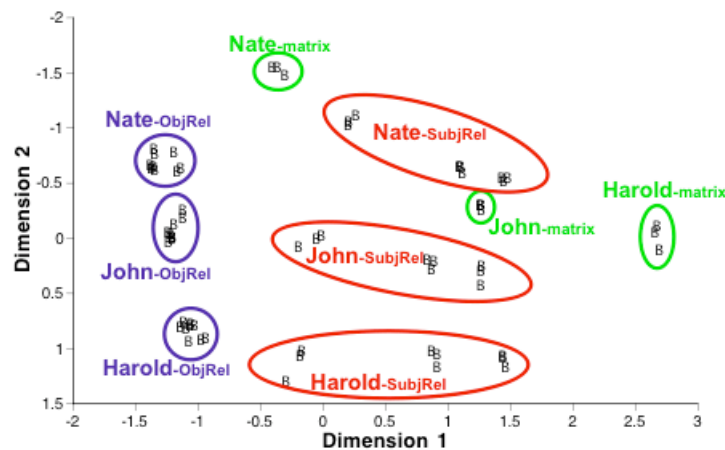


Figure 5: Multidimensional scaling results of network activation of different sentence types immediately prior to reflexive

along pairs of these axes. These methods have also been applied to the analysis of connectionist networks (Botvinick and Plaut 2004; Elman 1991, 1993, 1995; Rohde 2002).

We applied Multidimensional Scaling to the set of points representing the 63 sentence contexts described above, mapping them into a 2-dimensional space. (PCA yielded similar results.) The MDS plot, shown in Figure 5 revealed a feature of the representational space that was not evident in the HCA result: The 63 sentences cluster rather well into 15 groups, each group corresponding to a conjunction of features; 6 are conjunctions of subject and sentence type, and 9 are three-way conjunctions of subject, sentence type and name in the relative clause (in the case of subject relative). 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.

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 et al. 1995, 1996; Botvinick and Plaut 2004; Allen and 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. However, for sentence 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 phase-2 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 post-verb 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

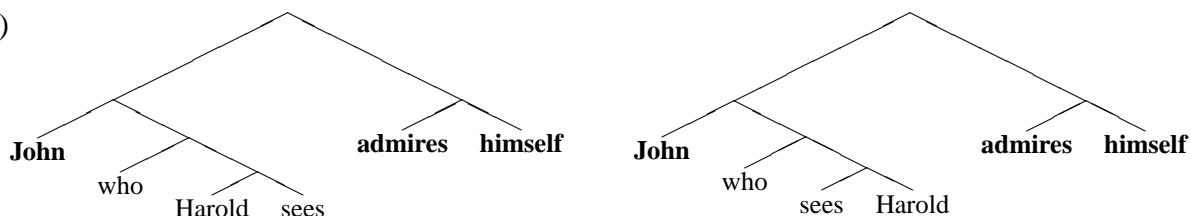
has not learned the abstract structural dependency between pronouns and their referents that was present in the target grammar.

## 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 networks internal structure does not faithfully mirror that of the target grammar. Yet it does not decisively demonstrate that in spite of these representational differences there is not some as yet undetected 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.

(19)



Independent of issues of anaphoric reference, the pronoun 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 two interpretive network using corpora that were derived from the one used in Experiment 1. The training corpus for No-SR-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-OR-Net was trained with the same corpus, but with 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 two 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.

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 precisely what Lewis and Elman (2001) found in their study of subject-auxiliary inversion, where the network 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

The No-SR-Net achieved 89% accuracy in reflexive interpretation on a stochastically generated test set of sentences not including subject relatives. In contrast, accuracy on a test set of sentences all of which included subject relatives was 32%. The No-OR-Net exhibited a similar disparity in performance.

## 4.2 Discussion

Our immediate conclusion from the results of this experiment indicate 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 with 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 was withheld: the sentences on which the No-SR-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-SR-Net achieves approximately 90% accuracy on reflexive interpretation for object relatives in comparison to approximately 30% accuracy for subject relatives (essentially chance over the gender appropriate interpretive options). The No-OR-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 this such overfitting is occurring in our case, we plotted the performance of the No-SR-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 never exceeds chance for gender matched interpretations during 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.

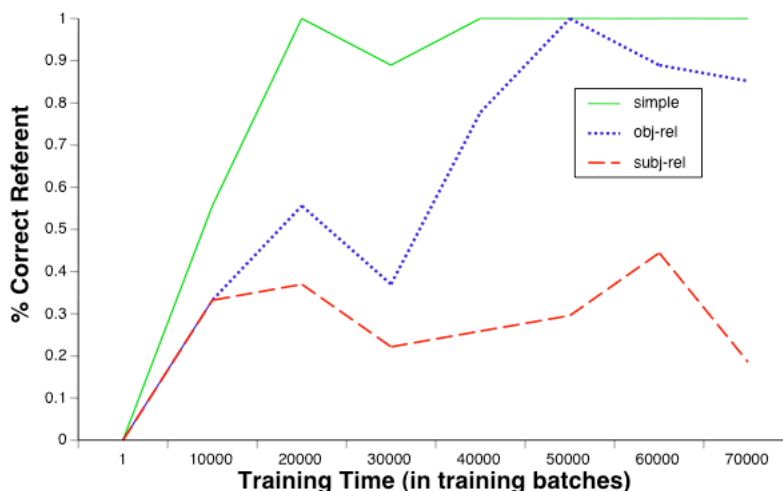


Figure 6: Performance of No-SR-Net on reflexive interpretation during training

## 5 Experiment 3: Lexical generalization

One question left unanswered by our previous experiments concerns another kind of variable in the rule that maps reflexives onto their antecedents. Specifically, the generalization governing reflexive interpretation is independent of the specific antecedent involved, but depends only on the structural context in which the reflexive is found. In this experiment, we investigated whether an SRN will induce a interpretive mapping for reflexive interpretation that generalizes across names, or whether it forms separate generalizations for distinct name-reflexive combinations.

To examine this issue, we again used the intact phase one network from Experiment 1, and re-trained the phase two network. As in the previous experiment, we used the training corpus from Experiment 1, but this time withheld training on the interpretation of reflexives whose interpretation was one of the referents in the domain (*John*). As in Experiment 2, it is not the case that sentences with *John* as antecedent of a reflexive never occurred during the training of the (phase two) network. Rather, the network was simply not given feedback during phase two training about the appropriateness of its interpretive output for reflexives in sentences of this sort.

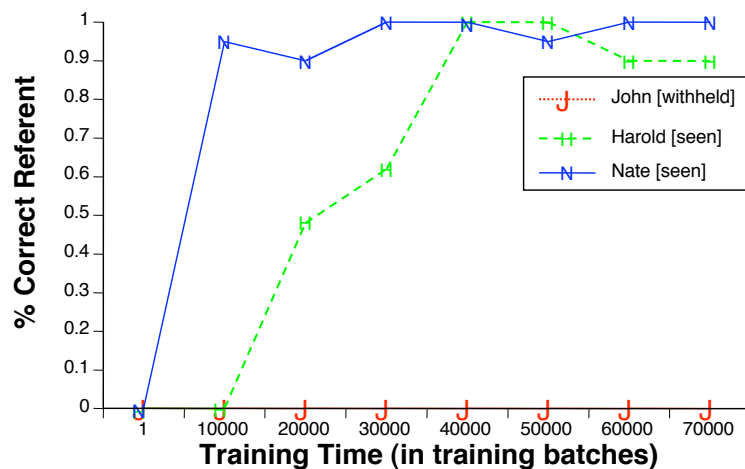


Figure 7: Performance on reflexive interpretation for withheld vs. non-withheld names during training

## 5.1 Results

For non-withheld names, the network achieved performance comparable to that in the previous studies. On a stochastically generated test set, the network assigned the correct referent to the reflexive in 91% of the cases. For the withheld name however, the network's performance was strikingly different. It never correctly assigned the correct referent for the reflexive, even in simple sentences such as *John saw himself*. Further, this pattern held throughout training, as Figure 7 shows.

This behavior recalls Marcus's (2001) "a rose is a rose" problem, but in a more natural cognitive domain. Marcus demonstrated that an SRN trained 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 novel nouns like *rose*. That is, after training, a network which had not been trained on the sentence *a rose is a rose* would not correctly predict the next word after receiving as input the sequence *a rose is a*. It seems to us reasonable to question the force of this argument on the basis of the following objection: Since the network has no experience whatsoever with the noun *rose* and because the 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. Notice however that because of our two phase training design, this objection does not apply to the design in our current experiment. During phase one training, the network has acquired extensive knowledge of the name for which reflexive interpretation is later withheld. Therefore, the input to the phase two network is a distributed representation that has been induced by the network and should reflect commonalities across the lexical items. If the interpretive network from phase two had developed a rule-like generalization that applied to all names, where that term is defined in terms of the common properties names share in activation levels in the phase one network's context layer, then the network should have been able to correctly resolve the interpretation of the reflexive even in the case of the withheld name. It is of course possible that a name-independent representation of reflexive interpretation can be represented in an SRN and induced under other conditions of training. However it remains to be shown just what such conditions would be and whether they plausibly match the context of child language acquisition.

## 6 Conclusion

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 Experiment 1 suggest that SRNs are biased to induce string-based dependencies even when such linear dependencies do not supply reliable cues to interpretation in the training data.

Moreover, we also found that the network's approach to anaphoric interpretation did not exploit variable-based generalizations that cut across different structures and different referents. The analysis of Experiment 1 coupled with the results of Experiments 2 and 3 point to the fact that SRNs tend 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 seem unable to extend the interpretive mapping for reflexives 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. Nonetheless, there have been a number of proposals in the literature on acquisition that suggest that learning of what come to be abstract grammatical generalizations start off as tied to specific lexical items and constructions Goldberg (1998); Tomasello (1992, 2000). Therefore the results of the experiments reported here 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, or whether it demands a different approach to grammatical knowledge in the form of an algebraic system of rules and representation.

## References

- Allen, J. and Seidenberg, M. S. (1999). The emergence of grammaticality in connectionist nets. In MacWhinney, B., editor, *The Emergence of Language*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Asudeh, A. and Keller, F. (2001). Experimental evidence for a predication-based binding theory. In Andronis, M., Ball, C., Elston, H., and Neuvel, S., editors, *Papers from the 37th Annual Meeting of the Chicago Linguistic Society*, volume 1, pages 1–14, Chicago.
- Badecker, W. and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphora. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28:748–769.
- Botvinick, M. and Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connections approach to normal and impaired routine sequential action. *Psychological Review*, 111:395–429.
- Crain, S. and Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 63(3):522–543.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.
- Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, pages 195–225.
- Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99.
- Elman, J. L. (1995). Language as a dynamical system. In Port, R. and van Gelder, T., editors, *Mind as Motion: Explorations in the Dynamics of Cognition*, pages 195–236. MIT Press, Cambridge, Mass.
- Elman, J. L. (1998a). Generalization, simple recurrent networks, and the emergence of structure. In Gernsbacher, M. and Derry, S., editors, *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Lawrence Erlbaum Associates.
- Elman, J. L. (1998b). Representational issues: Commentary on *the algebraic mind*. Unpublished manuscript, UC San Diego.

- Elman, J. L. (2002). Generalization from sparse input. In Andronis, M., Debenport, E., Pycha, A., and Yoshimura, K., editors, *Proceedings of the 38th Annual Meeting of the Chicago Linguistics Society, Volume 2*, pages 175–200.
- Elman, J. L., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness*. MIT Press, Cambridge, MA.
- Fisher, C. (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.
- Franck, J., Vigliocco, G., and Nicol, J. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, 17(4):371–404.
- Goldberg, A. E. (1998). Patterns of experience in patterns of language. In Tomasello, M., editor, *The New Psychology of Language*, pages 203–219. Lawrence Erlbaum Associates, Mahwah, NJ.
- Gordon, P. C. and Hendrick, R. (1997). Intuitive knowledge of linguistic co-reference. *Cognition*, 62:325–370.
- Haskell, T. and MacDonald, M. (2003). Proximity does matter: Evidence for distributional effects in the production of subject-verb agreement. Manuscript, USC.
- Hemforth, B. and Konieczny, L. (2003). Proximity in agreement errors. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Lawrence Erlbaum Associates.
- Joanisse, M. and Seidenberg, M. (2003). Phonology and syntax in specific language impairment: Evidence from a connectionist model. *Brain and Language*, 86:40–56.
- Koster, J. and Reuland, E., editors (1991). *Long-Distance Anaphora*. Cambridge University Press, Cambridge.
- Lewis, J. D. and Elman, J. L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th Annual Boston University Conference on Language Development*.
- Marcus, G. (2001). *The Algebraic Mind*. MIT Press, Cambridge, MA.

- Plaut, D. C., McClelland, J. L., and Seidenberg, M. S. (1995). Reading exception words and pseudowords: Are two routes really necessary? In Levy, J., Bairaktaris, D., Bullinaria, J., and P.Cairns, editors, *Connectionist Models of Memory and Language*, pages 145–159. UCL Press, London.
- Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: Computation principles in quasi-regular domains. *Psychological Review*, 103:56–115.
- Rodriguez, P. (2001). Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation*, 13(9):2093–2118.
- Rodriguez, P., Wiles, J., and Elman, J. L. (1999). A recurrent neural network that learns to count. *Connection Science*, 11:5–40.
- Rohde, D. (1999a). A connectionist model of sentence comprehension and production. PhD thesis proposal, Carnegie Mellon University.
- Rohde, D. (1999b). The Simple Language Generator: Encoding complex languages with simple grammars. Technical Report CMU-CS-99-123, Carnegie Mellon University, Department of Computer Science.
- Rohde, D. (2002). *A Connectionist Model of Sentence Comprehension and Production*. PhD thesis, Carnegie Mellon University.
- Rohde, D. and Plaut, D. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72:67–109.
- Rumelhart, D. E. and McClelland, J. L. (1986). On learning the past tenses of English verbs. In McClelland, J. L. and Rumelhart, D. E., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2, chapter 18. MIT Press, Cambridge, MA.
- Runner, J. T., Sussman, R. S., and Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: evidence from eye movements. *Cognition*, 89:B1–B13.
- Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48:542–562.

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

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