Paper version of an oral presentation, Symposium on Cognitive Architecture. In M.A. Gernsbacher & S. Derry (Eds.)., *Proceedings of the 20th Annual Conference of the Cognitive Science Society.* Mahway, NJ: Lawrence Erlbaum Associates. 1998.

Generalization, simple recurrent networks, and the emergence of structure

Jeff Elman

Department of Cognitive Science
University of California, San Diego
elman@crl.ucsd.edu

Introduction

If human behavior were list-like, accounting for human behavior would be simple: Just enumerate the list of possible stereotypies. Alternatively, if behavior were predictable on the basis of abstract, fully-productive, context-insensitive rules, our task would be different but similarly straightforward: just list the underlying rules.

The problem is that most human behaviors seem to lie somewhere between these two possibilities. Neither lists nor rules capture the richness of cognitive behaviors, which are populated by what Plaut, McClelland, Seidenberg, and Patterson have called "quasi-regular" domains. There are underlying generalities to most of what we do (or think), but the generalities are typically partial and tempered by qualifications. The problem is not only in knowing when a generalization should apply, but when it should not.

Language is a domain which is particularly interesting from this perspective, and the overgeneralization of regularities by children who are learning language is often cited as evidence for the rule-like nature of language. However, although much has been made of the productivity of child language learners, it is also true that children are enormously conservative. Overgeneralizations are exceptional, or at least not as common as would would think given the vast literature on overgeneralizations. If there is any rule, it is that children do not make rules as readily as has been supposed.

Our problem in some sense is how to have it both ways: How do we account for productivity and generality, while also accounting for limiting effects of context, item-specific information, frequency, category structure, and so on—effects which themselves range from general and predictable to purely idiosyncratic?

To many of us, connectionist models have seemed to give us a good start at solving this problem. It is worth noting that we do not come to this position out of perversity. Many of us, myself included, began within the symbolic tradition. My own interest was motivated by frustration with generatively-based linguistic theories which seemed repeatedly to bump against the same problems, while ignoring many others.

But connectionist models have been criticized, almost since their beginning, for being inadequate to the task. One of the strongest criticisms is that they cannot generalize properly. Fodor and Pylyshyn (1988) argued that connectionist models cannot account for productive behavior or

support inferencing, because they do not support truly compositional representations. Pinker and Prince (1988) argued that the generalizations were incorrect (and sometimes too strong). Hadley (1992) distinguished various types of generalization, and argued that connectionist networks were capable only of what he termed weak systematicity.

Most recently, Gary Marcus (1998) has stated, "[current connectionist models] cannot generalize outside the training space."

This is an interesting statement, because the claim is both potentially quite damning but also ambiguous. That is, several things might be meant by "generalizing outside the training space." In particular, at least two very different things might be meant.

(1) The first is what I take to be the usually understood sense. A training set is understood to consist of a finite number of examples of input/out pairings (or associations, or function mappings, etc.). The domain (input space) and range (output space) are pre-defined, and generalization occurs when the correct output is produce for an input not encountered during training.

Thus, we might ask whether, if a network is only trained on instances of "John" or "boy" in subject position, the network can correctly process instances of "John" or "boy" which occur in object position. This is an important question (and analogous to one considered experimentally by child language researchers) and one I will return to in the second part of this paper.

(2) The second sense in which one could mean "generalization outside the training space" is a different and very narrow one. One could mean: A network, having been trained on inputs which are defined in terms of one set of perceptual features, will not know what to do when it encounters an input which is defined in terms of a novel set of perceptual features. In this case, if a network which had been trained on stimuli presented in the visible light spectrum were unable to deal with stimuli in the ultra-red region of the spectrum, this would constitute a second sort of "failure to generalize outside the training space.".

This is in fact the kind of generalization failure that Marcus intends.

In remainder of this paper I want to do two things. First, I wish to consider examples of this second category of generalization failure. I take such failures to be untroublesome and will suggest they are entirely appropriate in networks, and similar to failures we find in humans and other biological organisms. Although I do not regard this as a very interesting aspect of generalization, it is worth considering simply to avoid confusions about just what networks can and cannot do.

Then I will return to the first sort of generalization problem, in which a network must deal with the novel use of inputs; for example, processing words which are encountered in syntactic contexts not seen during training. I believe this is in fact a very important issue about which we know some things, but for which there are many important questions to be studied. I will report several recent simulations which study the conditions under which networks generalize, given scant or gappy data.

Generalization I: Perceiving novel features

When Marcus claims that networks do not generalize outside their training space, what he has in mind is primarily what could be called "perceptual generalization." Marcus's point is easy to illustrate, although it is also very easy to be confused about exactly what the significance of the point is.

First, The problem is *not* that the networks are unable to deal with novel inputs. Networks can, and do, all the time.

For instance, consider the XOR problem. This is a Boolean function in which 2-dimensional inputs of 0s and 1s are grouped into two categories according to the logical disjunction operator, XOR:

Input		Output
0	0	0
1	1	0
0	1	1
1	0	1

The four input patterns can be visualized as four points in 2-dimensional space (see Figure 1). Each dimension (*x* or *y*) is used to encode one of the two inputs, and is seen by the network as

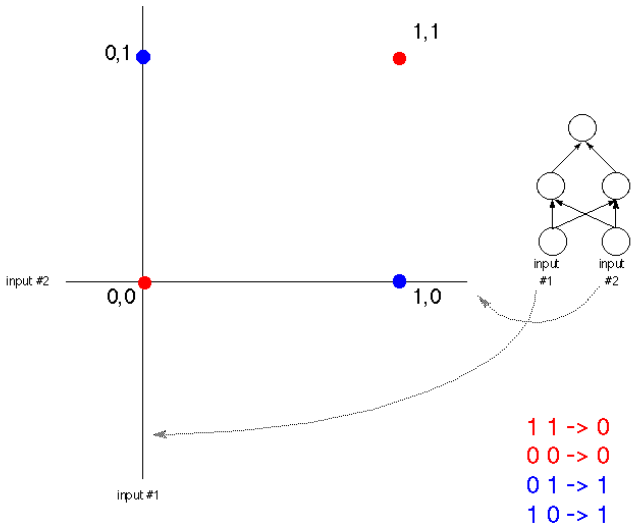
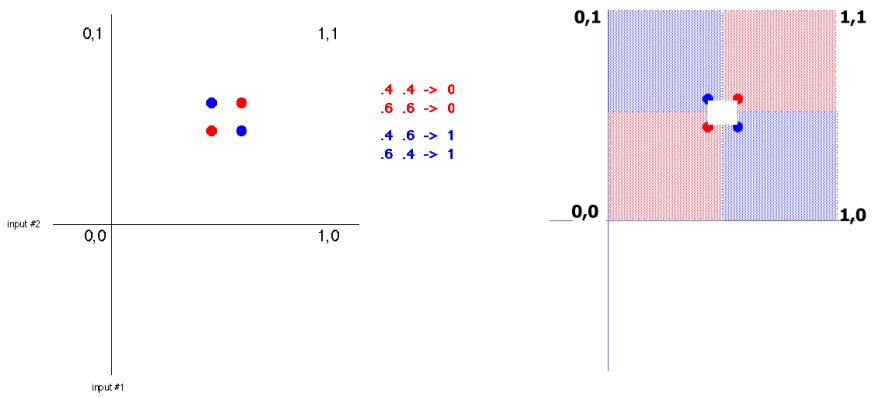


Figure 1.The four input patterns (00, 11, 01, 10) shown as points in 2-D space. Each input is encoded as a value along one dimension (x or y).

a value along one of the network's two input lines.

But we can train a network on another version of XOR in which the values .4 and .6 are used instead of 0 and 1; the logical relationship remains the same, although the training stimuli occupy a different region of the training space (Figure 2a).



(a) (b)

Figure 2.(a) The four input patterns for XOR are now represented as .4,.4; .6,.6; .4,.6; .6,.4. (b) The network will generalize after training on the four input patterns to an infinite number of patterns in the quadrant around each pattern.

If we do train this way, network will generalize to other points in the quadrant (although typically, not within the smaller region contained within ({.4,.4}, {.6,.6}}), that is, to an infinite number of stimuli (Figure 2b). So networks can generalize to stimuli it has not seen during training.

But Marcus knows this, and this is not what he is concerned about. What he points out is that such a network, cannot deal with a novel class of inputs containing, in this case, three input values (see Figure 3).

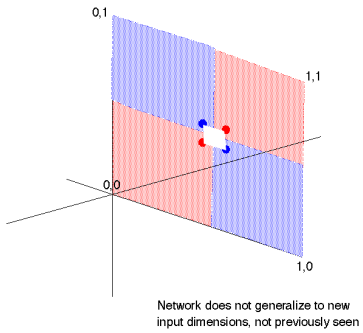


Figure 3. If a network trained on 2-D inputs is subsequently presented with test inputs which involve a third dimension, it has no way to generalize into this new input space—nor should it.

This is perfectly true. Moreover, it does not help to train with an extra dimension (or input node) from the start. If that dimension is not used, the weight connecting the input to the hidden layer will tend toward 0, and the input dimension will not longer be "seen."

This is a well-known consequence of back-propagation learning. What is interesting is that while Marcus takes this to be a crucial flaw in network learning, a more reasonable interpretation is that the network is doing exactly what humans and other animals do. Consider, for example, the visual system of cats. Blakemore and Cooper (1970) demonstrated many years ago that systematically depriving a cat of exposure to horizontal stripes during early life resulted in failure of neurons to develop which were sensitive to horizontal stripes.

Figure 4. Blakemore & Cooper (1970) demonstrated that a cat, reared during its early life in an environment which contained no horizontal stripes, would fail to develop neurons in the visual system which respond to horizontal edges.

Similarly, young human infants appear to be sensitive to the full range of speech sounds found in all human languages, but at older ages discriminate only those phonemic contrasts found in their environment. The difficulty of Japanese speakers to discriminate r/l or of English speakers to discriminate prevoiced from voiceless unaspirated bilabials is well-known.

The common result is essentially that, if perceptual experience is limited—either by evolution or learning—one will not be able to perceive things outside that experience. We do not perceive in the infrared, although pit vipers do.

Keeping this in mind, it is worth pointing out that human infants have a far richer perceptual experience than do networks. So it is very easy to run an experiment in which infants appear to generalize to novel stimuli, in cases where networks will not. In fact, those stimuli may be novel within the experiment but are composed of perceptual features with which the infant has ample experience outside the experiment. The perceptual experience of networks on the other hand, is typically (though it need not be) limited to the simulation.

This problem will be especially severe for networks which use localist representations, in which one perceptual dimension is assigned to every different entity. In such cases the network really does have to see all possible inputs in order to be able to respond at all to them. Marcus is quite keen on this, especially since many simulations of language processing have employed localist representations.

Why are localist representations used in the first place? There are at least two reasons why modelers may choose to use localist representations.

(1) First, localist representations of *outputs* are useful from the point of view of the modeler *qua* network analyst. Because localist representations assign each possible output response to a separate unit, the activations of the outputs can be interpreted (for certain tasks) as a probability estimate. This can provide us with a useful measure of the performance, especially if we wish to compare the network's performance with other probability-based models..

(2) Second, localist representations of *inputs* make the learning problem deliberately harder for networks because they lack any form-based similarity structure. The 1-in-n encoding of localist representations yields a set of orthogonal vectors, each one equally similar (i.e., equidistant in Euclidean space) to all others. Thus there is no information to be gleaned from the form of the inputs about how they are to be treated, or how they might be functionally similar to one another. Again, the use of localist representations represents a choice on the part of the modeler, reflecting strategic goals (e.g., to make learning deliberately difficult) rather than being dictated by necessity.

Are localist representations realistic? Clearly not. The complex entities we perceive in the real world have rich featural descriptions. The overlap and similarity structure of these featural representations play an important role in supporting generalization (although purely form-based similarity is not enough to get along in the world; tigers and cats may look similar, but don't try to keep a tiger as a house pet). Localist representations are also tremendously costly. Each entity requires a perceptual dimension for itself. This is as unrealistic and as costly as if natural languages were to assign a specific frequency in the sound spectrum for each unique word in our language. Given the resolving power of our auditory systems, we would end up with very small vocabularies.

Are localist representations necessary? Not at all. In Elman (1990) a simulation was described in which a network, trained on simple sentences of words that were encoded in a localist fashion, could extract the lexicogrammatical category structure underlying the grammar, as seen by the ways in which the hidden unit representations clustered. But it is possible to run the same simulation using distributed representations and get similar results (Figure 5).

The bottom lines are (1) localist representations are useful but not necessary to the connectionist models Marcus is concerned about; and (2) it is obviously true that distributed representations are more realistic; they afford a richer medium for representing the world, in which form-based similarity ("what you look like") can be used along with function-based similarity ("who you hang out with") to motivate generalization.

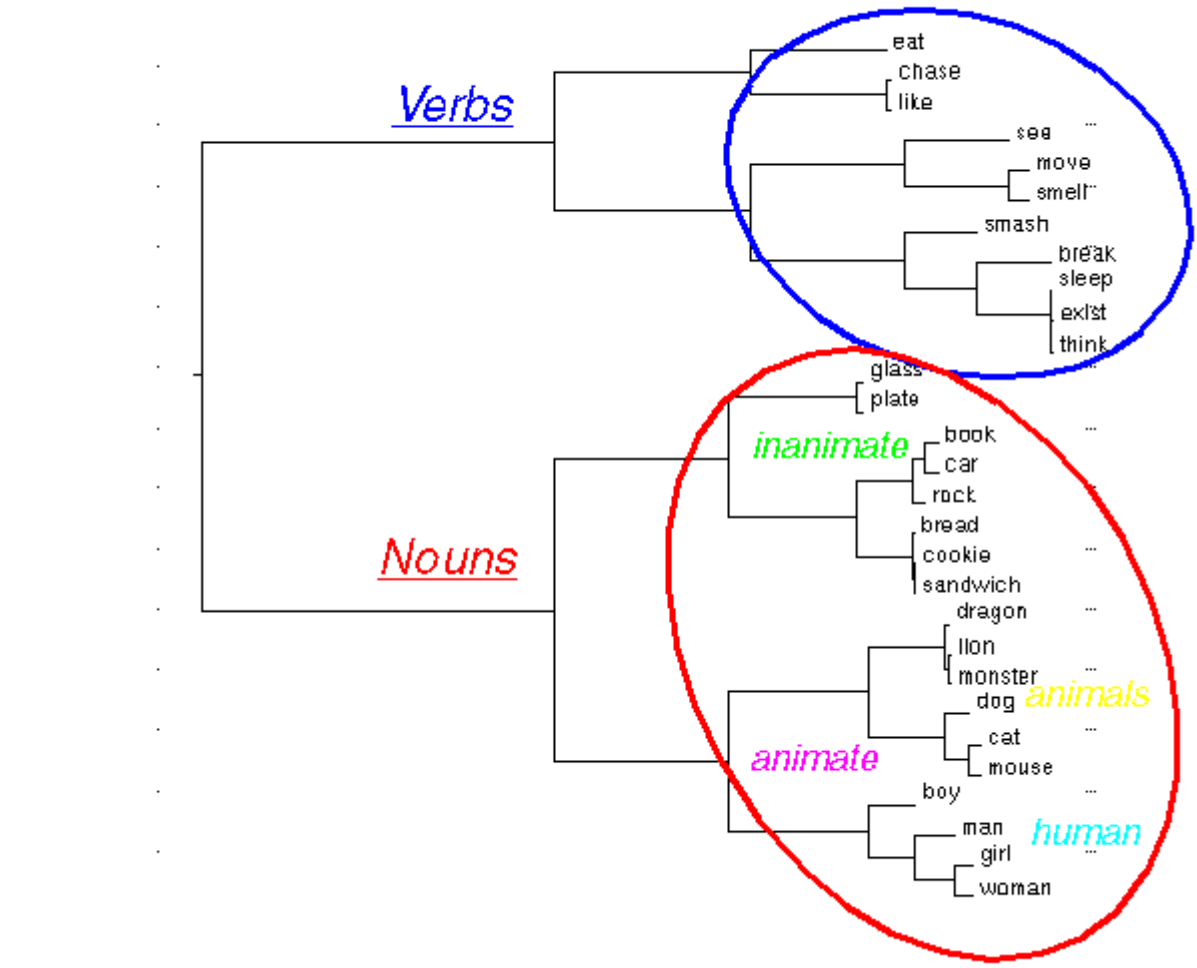


Figure 5. Hierarchical cluster of hidden unit activations from a network which was trained on simple sentences, with each word encoded as randomly chosen distributed representation.

Generalization II: Using words in novel contexts

I would like to turn now to the first sense of "generalizing outside the training space", because I think this raises more interesting issues.

The question can be posed, in specific form, thus: Can a network, trained only on sentences in which a given noun (e.g., "boy") appears only in subject position, deal appropriately with novel sentences in which that word appears in another syntactic context, e.g., object position. This sort of generalization is an instance of what Hadley (1992) has called "strong systematicity", and it has been claimed that simple recurrent networks are not capable of this sort of generalization (Hadley, 1992; Marcus, 1998).

The question is an important one, because it is very likely that the natural language input which children (and even adults) hear is extremely limited in just this sort of way. Although it is

definitely the case that children's linguistic input is both very rich and very extensive (Hart & Risley, 1995, estimate that by age 3, children hear between 10 and 30 million words), it is also almost certainly not the case that children hear all possible words in all possible syntactic contexts. Indeed, as one's vocabulary increases, the probability of encountering many words only a few times and in limited syntactic contexts increases.

Thus, the input from which we learn is both very extensive but also very gappy. If networks cannot extrapolate the appropriate use of a word to a novel context from limited exposure in the way that humans do, we (modelers) are in serious trouble.

As it turns out, networks *do* generalize. However, the conditions which enable such generalization are interesting, and remind us that the phenomenon of generalization across gaps in the input is not quite as simple as has sometimes been implied in the discussion of systematicity and productivity. This is illustrated in the following simulation (which, by the way, used localist representations, just to make the point that for this issue nothing critically hinges on localist vs. distributed representations).

Simulation

The network's task was to process a string of sentences, one word at a time, and predict successive words (e.g., Elman, 1990). Words differed with regard to their frequency of occurrence in the grammar, and verbs differed with regard to possible arguments, as well as preferred arguments. There were a number of simple grammatical constructions, several of which are shown in Figure 6 (font sizes indicate probability of occurrence in the corpus). A set of corpora were

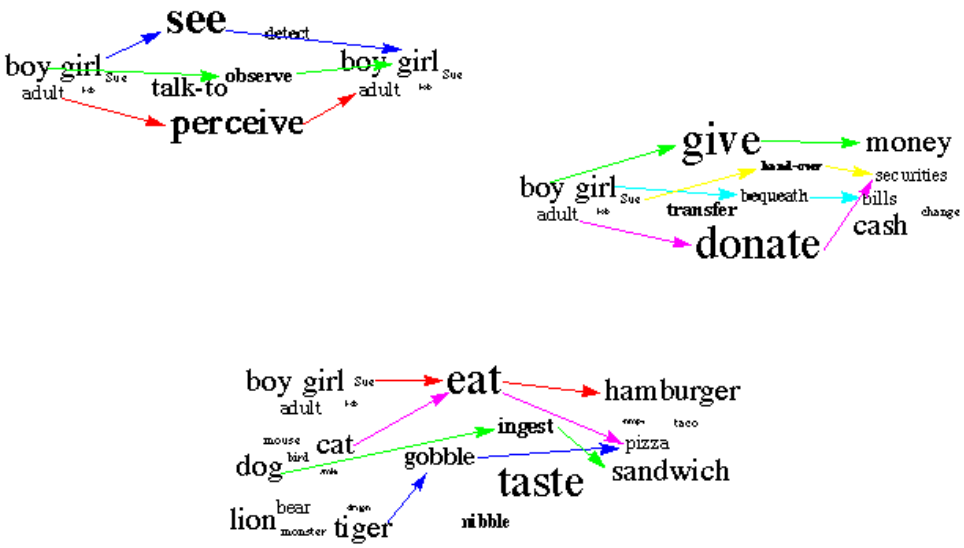


Figure 6. Schematic representation of some of the grammatical constructions used to generate a training database. Colored lines connect words which form sample sentences; font size indicates probability of occurrence in the corpus.

constructed consisting of random sentences from this grammar and ranging in size from very small (20 sentences) to medium (1,000 sentences) to large (5,000).

Although there were only 1,030 different possible sentences in this language, the low frequencies of some of the nouns mean that not all possible sentences occurred, even in reasonably large samplings (e.g. 5,000 sentences). In fact, a random sample of approximately a half million sentences is needed in order to ensure that all sentences are likely to appear.

Thus the data on which the network was trained are very gappy, which is probably a very realistic approximation of the situation in which children find themselves.

In deliberately constructing a training set for the network which is gappy, we are able to ask under what conditions the network generalizes across gaps (or doesn't). For example, in the network's artificial language, verbs of communication require human agents and direct objects. Thus, any of set of words "girl", "boy", "adult", "Sue", or "Bob" must serve as the agent and direct object of the verb "talk-to." However, with small corpora, particularly given that not all of these words occur equally often, it is possible that one or more may never appear in the training set as either agent of direct object. In fact, it was easy to find a corpus of 1,000 sentences in which "boy" never appears at all in direct object position *for any verb*. The question is then, never having seen "boy" as a direct object (although it does occur in the corpus as agent), will the network be unable to predict "boy" as a possible direct object following the verb "talk-to"?

Figure 7 shows the networks' behavior at three points in time: After 1,000 training trails, after 5,000 training trials, and after 10,000 training trials. During early learning (1,000 and 5,000 trials) the networks' expectations conform fairly closely to the raw statistics of the data. Given the sentence fragment "Girl talks to…" the network predicts a human direct object much more than it does the lexical item "boy." During these early stages of learning, the network can be said to be operating more or less with a rote strategy (cf. Plunkett & Marchman, 1993, for a similar pattern of learning of the past tense). However, at 10,000 training trials, the networks' predictions change markedly: Now "boy" is predicted as a grammatically acceptable item in direct object position—despite the fact that this word never occurs in that position.

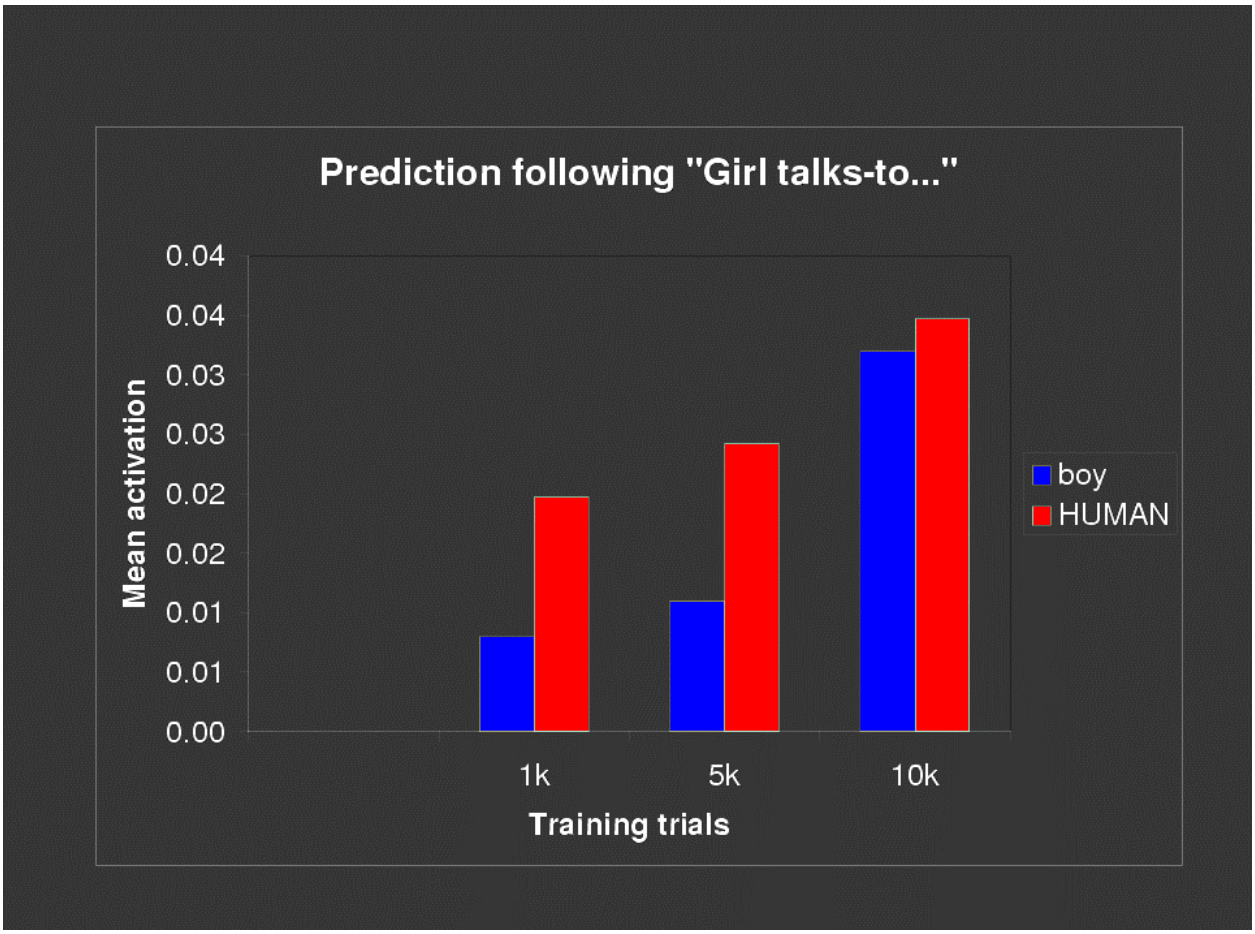


Figure 7. Network's predictions of "boy", compared with mean activations for other *human* nouns and for verbs, in the context "the girl talks to. . .".

Why does this occur? The answer is fairly straightforward.

In this simulation, the network sees only a fraction of the possible sentences. But importantly, although "boy" is never seen in direct object position, it is seen in shared contexts with other human words. For example, humans (but not other animals, food, etc.) appear as the agent of verbs such as "eat", "give", "transfer." Conversely, humans (including "boy") do not appear as agents of other verbs (e.g., "terrify", "chase", which in this language require animal agents). The word "boy" shares more in common with other human words than it does with non-human nouns, or with verbs.

In networks, as for human, similarity is a powerful motive force which can drive generalization. Similarity may be a matter form ("who you look like") or behavior ("who you hang out with"). In this simulation, words were encoded with localist representations, so there was no form-based similarity. But as we have just seen, there were behavior-based similarities between "boy" and other "humans."

These more abstract similarities are typically captured in the internal representations that network construct on their hidden layers, and they are what facilitate generalization. The overall

behaviors which “boy” shares with “girl”, “Sue”, “Bob”, etc., are sufficient to cause the network to develop an internal representation for “boy” which closely resembles that of the other human words. Figure 8 shows a hierarchical clustering of the internal representations of the lexical items known to the network, after 10,000 training trials. The similar of the internal representation of “boy” to representations of other human nouns is indicated by the proximity in hidden unit space of “boy” to those other items.

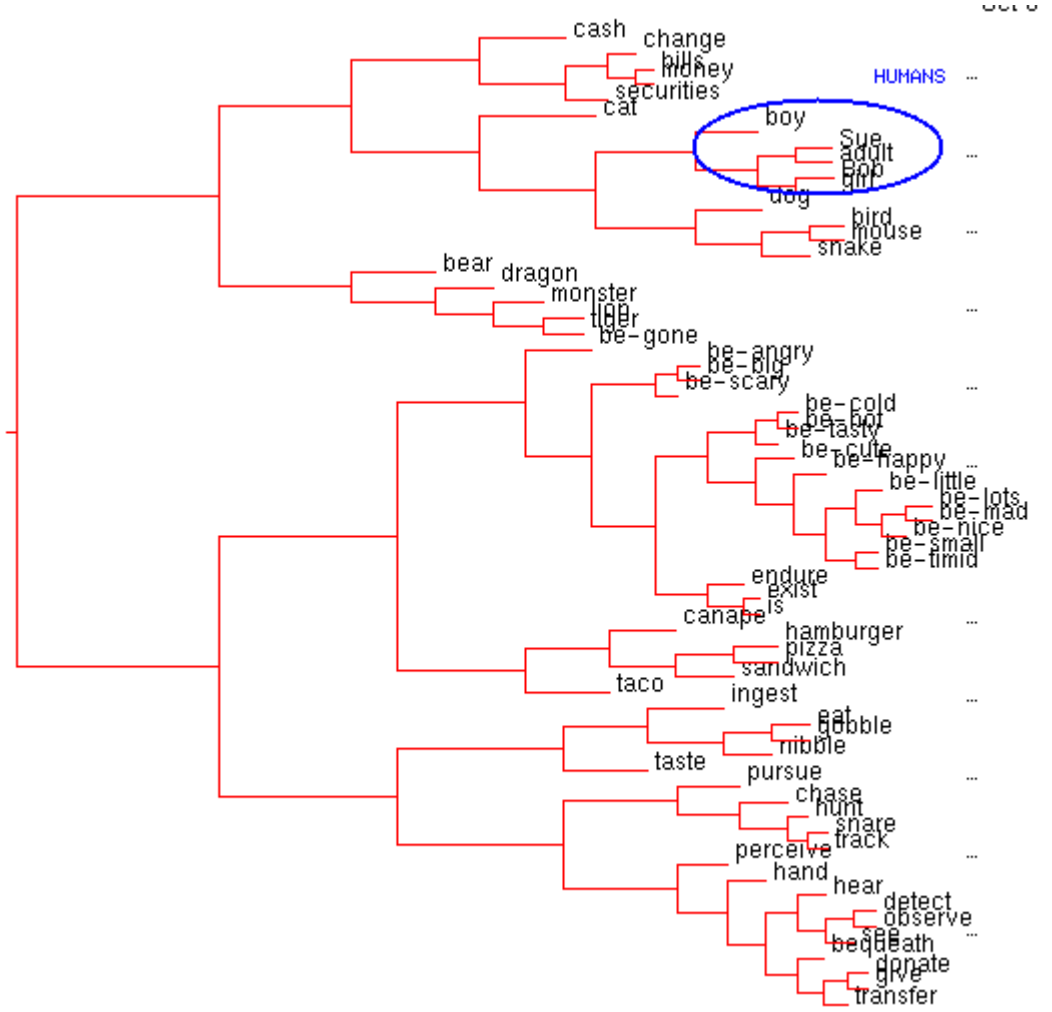


Figure 8. Hierarchical clustering of the internal representations (hidden unit activation patterns) of the lexical items known to the network, after 10,000 training trials.

However, the internal representations for those other words must reflect the possibility of appearing in direct object position following communication verbs (since the network does see many of them occurring that position). Since the representation for “boy” is similar, “boy” inherits the same behavior. The network’s knowledge about what “boy” can do is very much affected by what other similar words can do.

Of course, if the examples are too scant such generalizations are not made. With limited experience, the pattern of interlocking relationship which motivates the abstract categories is not

revealed. This is very much in line with what Michael Tomasello (among others) has noted with regard to children's acquisition of categories and constructions. Categories such as "noun" and "verb" do not start out as primitives; rather, they are accreted over time. At intermediate stages different words may be more or less assimilated to what will become adult categories (Olguin & Tomasello, 1993; Tomasello & Olguin, 1993).

There is a flip side to this coin: *Sometimes gaps are intentional*. For instance, the fact that "ungrasp" is not a possible word (although "unclench" is just fine; see Li & MacWhinney, 1996 for a network simulation of how this can be learned), or that even though both "the ice melted" or "she melted the ice" are acceptable paraphrases, one can say only "the ice disappeared" and not "she disappeared the ice." In other words, some gaps are not accidental but systematic—even if exactly what is systematic about the gap is not obvious. Thus together with the problem of generalization we have the related problem of over-generalization. Will the network always generalize through gaps?

The answer is no. Such generalizations depend on the relative amount of data and experience which are available. If the word "boy" appears overall with low probability, but there are sufficient other examples to warrant the inference that "boy" has properties similar to other words, the network will generalize to "boy" what it knows about the other words. However, if "boy" is a frequently occurring item, except in one context, the network is less likely to infer that a gap is accidental. It is as if the network realizes that the gap is not due to incomplete data (because the word is very frequent) and so must be the result of a systematic property of the word.

This can be seen by allowing the network to continue training on a corpus in which "boy" is absent (see Figure 9).

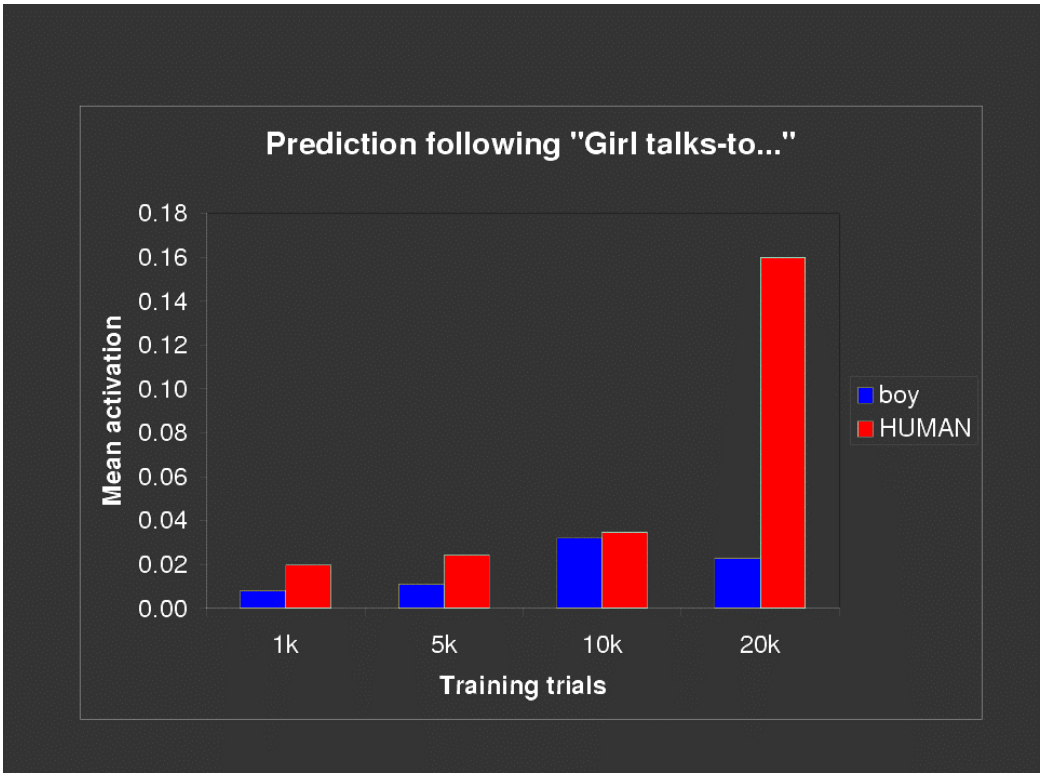


Figure 9. Network predictions at four stages in learning.

Although at 10,000 training trials the network generalizes the use of “boy” to object position, if there is significantly more training in which “boy” does not appear where it is now predicted the network retreats from that generalization. This occurs because of the indirect negative evidence provided when the network’s predictions are not confirmed by the data.

Thus the network goes through three stages. In early learning, the network hews close to the observed facts and is conservative in its predictions. With additional training, the network learns generalizations about classes of words, and this allows it to generalize to the novel use of familiar words. At this s tage, while the network’s experience is still somewhat limited, gaps are treated as accidental. However, with additional training, if the gaps persist the network learns to identify them as such. It is as if the network recognizes that it has now seen enough sentences that absences can no longer be attributed to small sample size and must be diagnostic of a systematic property of the word’s usage.

The lexicon and grammar

Finally, this simulation also reveals an interesting relationship between growth in the lexicon and the emergence of grammar, reminiscent of that noted by Bates and Goodman (1997). For the network, what it means to know a word is to know how to use it, i.e., its grammatical properties. As we’ve seen, this knowledge does not require exhaustive experience with a word in all possible contexts. As long as there is sufficient experience with other words which can collectively establish a category, the network will extrapolate to novel uses. If overall experience is too limited, and the network sees too few words or words in too few contexts, generalizations will not occur. Put another way, we might expect overall grammatical knowledge to increase to the corpus size, independently of experience.

Such a positive relationship between the size and richness of the lexicon and grammatical performance is shown in Figure 10. Twenty-two corpora were created; the smallest contained 20 sentences and the largest contained 5,000 sentences. Each corpus was used to train a different network. Each network was trained for the same period of time (80,000 sweeps) in order to hold training experience constant. Grammatical performance was measured by calculating how well a network’s empirical predictions about word usage compared with the actual possibilities, given the grammar (both actual and theoretical predictions can be represented as vectors; what is shown is the cosine, which measures the similarity between them).

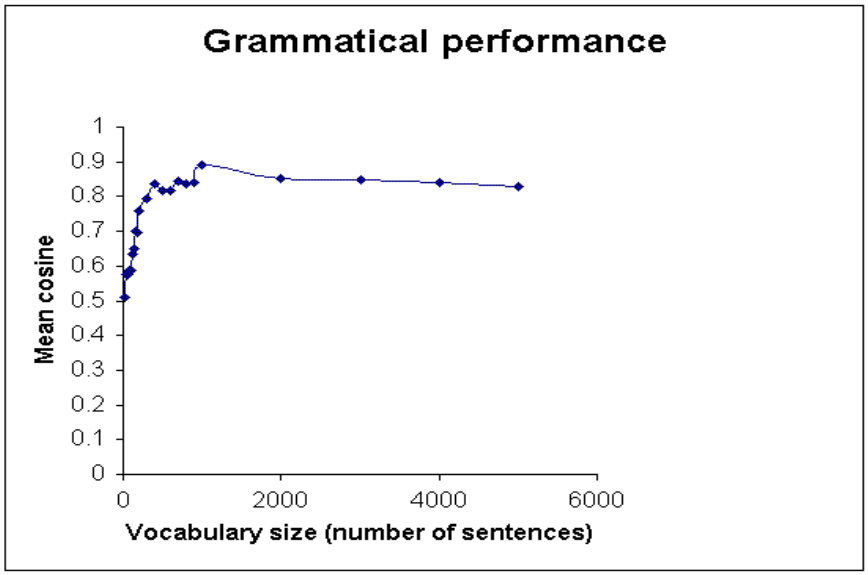


Figure 10. Increase in grammatical performance as a function of increasing corpus size, holding training sweeps constant. Grammaticality is measured as the cosine of two vectors: the networks' predictions of possible successor words, and the grammatically possible successors.

If performance were simply dependent on experience, one might have expected that networks with smaller-sized corpora would do better than the largest corpora. That's because, since all networks experienced the same amount of training, networks with smaller corpora have less to learn. The opposite pattern is found: Grammatical performance improves dramatically for networks that are trained on more sentences, even though each sentence is seen relatively fewer times. The improvement is greatest in the smaller range of corpus sizes and the rate of improvement rapidly drops.

The fall-off in rate of improvement is not because networks which have seen 1,000 or more sentences have essentially seen all sentences. They have not. A random sample of 1,000 sentences on average contains only about 30% of all possible sentences (recall that a half million sentences are needed to see the full set). Nor is the rapid rate of improvement at early stages due simply to the fact larger corpora allow networks to see all words in all contexts. This greater experience undoubtedly plays an important role, but it is also true that grammatical performance can increase in larger corpora even for words whose frequency of occurrence happens (by chance) to be no greater than in smaller corpora.

For example, in Figure 11 we see the change in grammatical performance for six words, as a function of occurring in different sized corpora. All six words improve as the corpus size increases. Yet the absolute frequency of occurrence for these six words was identical in all cor-

pora. Their improvement was due to their greater ability to piggyback off the what the networks with larger data sets were able to learn from other words in the same category.

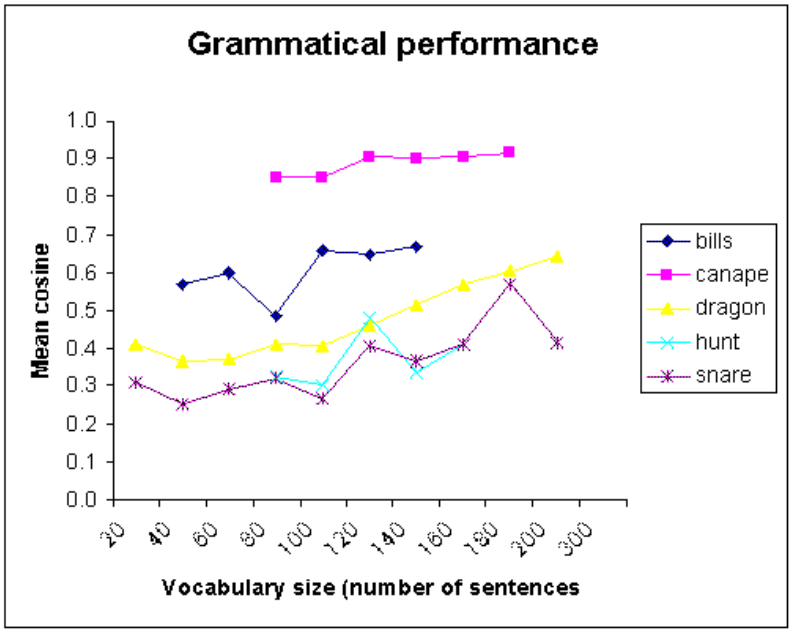


Figure 11. Increase in grammatical performance as a function of increasing corpus size, holding constant both training sweeps and also frequency of occurrence of each word in the various corpora.

This sort of relationship between lexicon and the grammar is just what Bates and Goodman (1997) have reported for children. It has long been known that most children undergo a rapid acceleration in the growth of their vocabulary, which usually occurs sometime between 16 and 20 months. What Bates and Goodman discovered is that there is a very tight relationship between this "vocabulary burst" and the emergence of grammar. When one of these developmental landmarks is delayed or accelerated, so too is the other. There are doubtless many factors which play a role in both vocabulary and grammatical growth, and which are not captured in the present simulation. However, the role of vocabulary size in generalization, and the role of generalization in supporting grammar, which is demonstrated in this simulation provides a plausible account for at least one of the factors underlying the vocabulary-grammar relationship. Details of this relationship are now being studied through further simulations.

Conclusion

I have tried to show that simple recurrent networks do in fact generalize to novel inputs and to novel uses of inputs. It is true that networks are subject—as are we—to perceptual limitations.

A network which is not designed to perceive along a given input dimension, or which is systematically deprived of experience in that dimension, will fare no better than a human who has not evolved to see in the infrared, or a cat who is reared in an environment containing no horizontal stripes.

But the more important observation is that the generalization process is complex, subtle, often partial, and rarely straightforward. There are critical effects of corpus size, corpus structure, and the time course of learning, and many open questions remain. For instance, not only is the size of a corpus and the frequency of exemplars important, but the way in which the data are structured can play a crucial role in categorization (e.g., Rodriguez, 1998). Some of the effects are counter-intuitive. But if the generalization process in networks is complex, subtle, often partial, and rarely straightforward, I believe this is also true of humans as well.

Acknowledgments

I thank Elizabeth Bates, Mary Hare, Kim Plunkett, and Paul Rodriguez for helpful their many helpful comments and suggestions at various stages of this work.

References

Bates, E., & Goodman, J.C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia, and real-time processing. *Language and Cognitive Processes*, 12, 507-584.

Blakemore, C., & Cooper, G. (1970). Development of the brain depends on the visual environment. *Nature*, 228, 477-478.

Elman, J.L. Finding structure in time. *Cognitive Science*, 14, 179-211.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3-71.

Hadley, R.F. (1992). Compositionality and systematicity in connectionist language learning. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates. Pp. 659-670.

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore: Paul Brookes Publishing.

Li, P., & MacWhinney, B. (1996). Cryptotype, overgeneralization and competition: A connectionist model of the learning of English reversible prefixes. *Connection Science*, 8, 3-30.

Marcus, G. (1998). Symposium on Cognitive Architecture: The algebraic mind. In M.A. Gernsbacher & S. Derry (Eds.)., *Proceedings of the 20th Annual Conference of the Cognitive Science Society.* Mahway, NJ: Lawrence Erlbaum Associates. P. 6.

Olguin, R., & Tomasello, M. (1993). Two-year-olds do not have a grammatical category of verb. *Cognitive Development*, 8, 245-272.

Plunkett, K., & Marchman, V. (1993). From rote learning to s ystem building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48, 21-59.

Rodriguez, P. (1998). Exploring gang effects by output node similarity in neural networks. In M.A. Gernsbacher & S. Derry (Eds.)., *Proceedings of the 20th Annual Conference of the Cognitive Science Society.* Mahway, NJ: Lawrence Erlbaum Associates. P. 1260.

Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, 8, 451-464