Where Does Systematicity Come From?
Effects of Training Corpus Structure and Attention on Systematic Generalization

Mark F. St. John
Department of Cognitive Science
University of California, San Diego
La Jolla, CA 92093-0515
stjohn@cogsci.ucsd.edu

Abstract

Human language and memory are only quasi-systematic. They are composed of context free (systematic) mappings, context sensitive mappings, and idiosyncrasies. Consequently, generalizations to novel stimuli may be systematic if they result from the context free mappings or may become "regularized" toward known stimuli if they result from the context sensitive mappings. Two factors that affect the *degree* of systematicity are the structure of the training corpus and the *amount* of attention or vigilance paid to the task. More systematic training corpora and more attention produce more systematic responses and fewer specific context sensitive regularizations. A simple PDP model is used to demonstrate these phenomena. A 3-layer feedforward network learns an auto-associative mapping. Untrained stimuli are tested to see if the model will respond with the systematic generalization or with a specific regularization by activating the output pattern for the nearest trained neighbor.

For the learning of most cognitive skills, generalization is critically important. Language learning is the paramount case. Based on learning from a finite corpus, children must generalize to the infinite set of sentences in their language. Fodor & Pylyshyn (1988) argue that this massive generalization ability is due to the nature of language: its systematicity. Systematicity is here defined as the quality that each word refers to the same concept regardless of context. A systematic language is composed of independent form-meaning pairs. Novel sentences and propositions can be made simply by recombining these independent pairs. So, if a child can understand "Ricky played a trick on Lucy," then she should be able to understand "Lucy played a trick on Ricky." To learn a systematic language is to learn the form-meaning pairs and a grammar for relating the structure of the forms to the structure of the concepts. How might such learning take place? If induction of such functions is even possible, what sorts of processing architectures and stimuli can induce them? If it is only possible to a degree, what are the limits and what factors modulate that limit?

In spite of the elegance of the systematicity characterization of language, many argue that real language is systematic only to a first approximation. At a closer inspection, many aspects of language are unsystematic. For example, Goldberg (1992) and Pinker (1989) identify a number of verbs that are unsystematic yet productive in the constructions they permit. For example, "Lucy told Ethel the news" is acceptable, but "Lucy whispered Ethel the news" is not, despite the semantic similarity of the verbs. Pinker (1989) has pointed out that there are discernible semantic subclasses of verbs that do or do not permit the ditransitive construction, and novel verbs in these subclasses are productive.

Similar points have been made with regard to forming the past tense of English verbs (Rumelhart and McClelland, 1986) and pronouncing regularly and irregularly spelled English words (Seidenberg and McClelland, 1989).

A complex relationship, such as English spelling to pronunciation or English sentences to meanings, can be characterized as a set of context free (systematic) mappings, context sensitive mappings, and idiosyncrasies. I will call such relationships *quasi-systematic*. Perhaps all levels of language are best characterized in this way.

One concern is how quasi-systematic relations are learned and how trained examples are processed. Does it require learning a systematic function and a set of restrictions, or can the complexity be learned all of a piece? A second concern, and the concern to be addressed here, is how generalization cases are processed. Specifically, given a novel example, will a context free, context sensitive, or idiosyncratic mapping be chosen to produce a response? What factors of training, processing, and cognitive architecture affect the response?

In a distributed processing model, such as a PDP model, processing depends on similarity: novel test instances are processed like the trained instances to which they are most similar. In sophisticated

models, all trained instances contribute to processing the novel instance according to their similarity. Systematic training materials produce systematic generalizations because each of the features of a test instance will be similar to the features of many training instances. Therefore, each feature will be familiar and well supported for processing and remembering. The use of a similarity metric during processing and generalization is well worked out in research on concept categorization and recognition (Medin & Schaffer, 1978; Shepard, 1987; Nosofsky, 1988).

Brousse and Smolensky (1989) examined the effects of training corpus systematicity on generalization in a PDP model. They showed that a highly systematic training corpus could produce generalizations to a huge test corpus that maintained that systematicity. They also showed that novel stimuli that violated strong context sensitive dependencies found in the training corpus would be misprocessed to better correspond with those dependences. For example, a network was trained to associate English 4-letter words with themselves through a narrow information channel (an autoencoder). Novel test patterns that violated English orthography were often "regularized" to fit English.

"Regularization" is used in the literature to describe the modification of test examples to fit previously learned regularities which correspond to what I am calling the context sensitive mappings. What is often missing in these discussions is the realization that there are other possible mappings, in particular, the context free, systematic mapping. The systematic mapping, in the spelling autoencoder would be veridical reproduction of each novel example, rather than regularization to known examples.

A similar phenomenon arises in story comprehension. During recall, details are partly remembered and partly reconstructed from background knowledge (See Graesser, 1981 for a review). Even though atypical details show better discrimination in recognition tests, reconstruction from experience and guessing produce better *overall* recall and recognition for typical details. According to systematicity, any novel story should be represented, stored, and recalled equally well. According to quasi-systematicity, however, the recognition and recall of novel stories whose details contradict experience will be regularized to better fit with experience.

This sort of regularization can also be seen *during* comprehension. Erickson and Mattson (1981) gave subjects questions like, "How many animals of each kind did Moses take on the arc?" Most subjects readily answered "two," despite their knowledge, when later questioned, that Noah was the correct Biblical figure. Subjects' knowledge of the arc story apparently overrode their knowledge of Moses. The meaning of Moses was regularized to the meaning of Noah.

What I aim to investigate is how quasi-systematicity in the training corpus affects when novel stimuli are generalized in a systematic way and when they are regularized.

Attention

A critical objection, or addendum, to these ideas is that processing bizarre sentences and answering trick questions can, in fact, be done. It just requires a little extra attention. People certainly can read and understand sentences like "Lucy whispered the news to Ethel." People also quickly gain immunity to trick questions once they become wary and begin to pay closer attention. People can also proofread with some success by paying close attention.

Erickson and Mattson (1981) found that reorganizing the questions to make the violations more prominent, led more subjects to notice the incongruity. Conversely, reduced attention to the text should decrease the notice of incongruities. In the extreme, speed readers spend little time and attention on the details of a text and may be plagued by misunderstandings if the text is difficult (cf. Just and Carpenter, 1987). Taken together, these studies suggest that a relative lack of attention will produce specific regularizations, while close attention will produce more systematic, and veridical, processing.

Attention in language processing can take several forms. *Selective attention* operates as a filter to a limited capacity processor. *Vigilance* operates as greater care and closer inspection of the stimuli -- greater processing power. The concern here is with the *vigilance* form of attention. In this scheme, processing is inherently somewhat noisy. Difficult stimuli will sometimes be processed incorrectly because of this noise and responded to as if it were similar but better known stimuli -- they will be regularized. Attention boosts the processing signal so that even these difficult stimuli may be processed correctly.

To demonstrate these points, let's consider a very simple case: the processing of simple binary patterns. Such simple patterns are far removed from the complexities of real language, perhaps most importantly, removed from language's hierarchical structure. Yet, this simple case allows several points to be made clearly.

Simulation 1 - Systematicity and context sensitive regularity

The task of the model is auto-association, that is, to reproduce an input pattern over a set of output units. This task can be thought of as a very simple version of many cognitive tasks such as comprehension from words to concepts, concept categorization, perception, or recognition memory (where the strength of the response is a measure of familiarity). The input/output patterns, and the correspond-

ing model, consist of two banks of units with one unit turned on in each bank. So, the pattern 1•1 corresponds to turning on the first unit in bank 1 and the first unit in bank 2. Pattern 1•2 corresponds to turning on the first unit in bank 1 and the second unit in bank 2. There are 100 such possible patterns.

Table 1
Training Stimuli

1•1						
2•1	2•2					
3•1	3•2	3•3	3•4			
4•1	4•2	4•3	4•4	4•5	...	4•10
5•1	5•2	5•3	5•4	5•5	...	5•10
...						
10•1	10•2	10•3	10•4	10•5	...	10•10

The model is trained on a subset of these 100 patterns, and then tested on the untrained patterns for its ability to generalize. The set of trained patterns is shown in table 1. The trade-off between systematic generalization and specific regularization is investigated by manipulating the number of similar training patterns and their systematicity. 1•10 should show poor systematicity but strong specific regularity because of the paucity of 1•X training instances. 2•10 and 3•10 should show increasingly better systematicity because of the increasing number of systematic neighbors.

The architecture is a feedforward auto-encoder. Both banks of input units are fully connected to the hidden layer, and the hidden layer is fully connected to the output layer (See figure 1). The hidden layer contained 15 units. The model was trained on the corpus in table 1 for 2000 epochs. The learning rate was .01 and momentum was .9. Weights were changed after each full epoch.

Results The model learned the training stimuli perfectly. The question is how the model generalizes to the untrained cases. Performance was quite good. For the cases 1•2 through 1•10, the model activated 1•X, the systematic response, most strongly for 7 of the 9 cases, with an average activation of .79. Only one nonsystematic was made for all of the 2•X or 3•X cases. For illustration in table 2, the systematic responses were averaged and shown as X•10. The average nonsystematic response for each unit is also shown for each unit. The ratio of systematic activations to specific regularity activations increases from 1•X through 3•X: .79/.65 = 1.22, .83/.29 = 2.86, .60/.19 = 3.16.

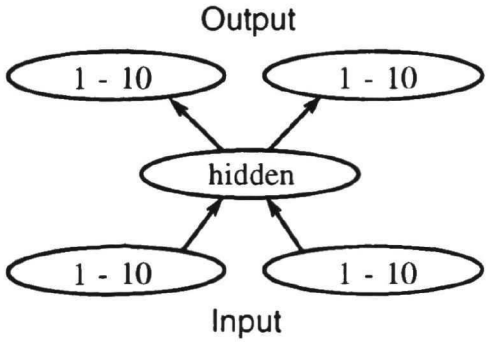

Figure 1. Architecture of the network.

Both nonsystematic responses in the 1•X cases involved greater activation of 1•1 than of 1•X. Even when the systematic response is stronger, the average activation of 1•1 is rather high. These "ghosts" of trained stimuli are smaller for 2•X and 3•X because of the greater systematicity among their training cases. Ghosts show up even more strongly if we present just a partial pattern to the input. The model completes the pattern with the ghosts of the trained stimuli. These ghosts, of course, represent the effects of the specific, context sensitive, regularities.

Table 2
Output Activations to Novel Stimuli

pattern	1	2	3	4	5	6	7	8	9	10
1•10	.65	0	0	0	0	0	0	0	0	.79
2•10	.20	.09	0	0	0	0	0	0	0	.83
3•10	.05	.05	.07	.02	0	0	0	0	0	.60
4•10	0	0	0	0	0	0	0	0	0	.96

The effect of the *size* of the training corpus on systematicity was also tested. A fresh network was trained on the subset of the training corpus created by removing the 6•X through 10•X stimuli (See table 1.). The model produced the systematic generalizations on 67% of the generalization cases, compared to 87% when trained on the original corpus. The corpus can be reduced further by removing X•6 through X•10. A fresh network trained on this smallest corpus of 17 stimuli generalized systematically on only 50% of the test cases.

Discussion These data demonstrate three points. First, with the largest most systematic corpus, the trained model produces systematic generalizations for nearly every test case: 20 out of 23 test cases. This strongly systematic performance depends on the systematicity of the training corpus. Smaller, less systematic corpora reduced the systematic generalizations and increased the specific context sensitive regularization generalizations.

Second, the nonsystematic responses which the model produces are not random. Rather, they fit specific regularities found in the corpus. For example, the response to 1•9 is 1•1, which was the only 1•X training instance. These specific regularities show up strongly in pattern completion, or "inference," tests. In a sense, these specific regularity ghosts are always lurking in the background, ready to appear when the systematic response is

weak or absent.

These results fit with the concept of identity and associative constraints proposed by St. John (1992). Constraints pertain to the mapping from input to output patterns. Identity constraints are those constraints or mappings that specify a one to one mapping between input and output patterns or between forms and meanings: "Lucy" means *Lucy*, and in the current simulations, 1 means 1, and so on. Identity constraints provide the building blocks for systematicity by specifying the independent form-meaning pairs. Associative constraints are all the other constraints. They encode regularities between pairs, and they are useful for drawing inferences and pattern completion.

Both identity and associative constraints are always learned, but their relative strength depends upon the structure of the whole training corpus. Systematic corpora produce strong identity constraints because identity constraints are the only constraints that hold between the input and output. For example, for the pattern 10•10, each integer predicts itself, but there is no predictability between integers. To process this instance correctly, the model *must* learn the identity constraints. On the other hand, the associative constraint from 1• to •1 is strong because given 1•, •1 is perfectly predictable. The associative constraints from 2• to •1 and •2 are each less strong because given 2•, each output pattern is only partially predictable.

Third, the model's responses are not all-or-none systematicity or regularization. Instead, responses are *graded*. When the identity constraints are relatively strong and the test item is similar to a number of trained items, as is the case for the 3•X items, the activation of units which correspond to the systematic response are strong, and the ghost activations which correspond to the specific regularities are very weak. Conversely, when the identity constraints are relatively weak and the test item is similar to few trained items, as is the case for the 1•X items, responses are only moderately systematic and the specific regularities are moderately strong.

Simulation 2 - Attention

How does vigilance attention affect processing? Can it improve systematic generalization? Our intuition is that it can. But first, we need to consider how responses are actually made. In the previous simulations, responses were reported as activation levels. These activation levels can be converted into response probabilities according to the Luce choice rule (Luce, 1963; McClelland, 1991). The probability of any response is set to the ratio of activation of the unit corresponding to that response divided by the sum of the activations of all responses.

An alternative is to make processing in the network itself probabilistic (McClelland, 1991). One method that McClelland suggests is to add noise to the input signal. The idea is that preprocessing of the stimulus itself produces noise.

Next, the input signal is modulated to implement attention. Under normal processing, the noisy input signal is attenuated. Under close attention, the noisy input signal is magnified. The activations of input units are set to be

$$(1) \qquad A_i = (I_i + \alpha * \text{noise}) * \text{attention}$$

where I_i is the value of the input pattern (0 or 1), noise is a random number normally distributed between -1 and 1, α is a parameter that determines the magnitude of the noise, and attention is a parameter that determines the magnitude of attention. Processing then proceeds deterministically as normal, but the strongest output activation is now taken to be the actual response. The Luce choice rule is not applied.

Attention modulates the sum of the input signal plus noise because it is assumed here that the input to the model is already corrupted by noise. There is no means for attention to preferentially weigh the signal. Instead, its job is to maximize performance given an already corrupted input.

Results The network was trained on the full corpus in table 1 *without noise* and with the attention parameter set to 1.0. For testing the generalization cases, α in (1) was set to 0.5. Under low attention (*attention* = 0.5), for 1•X, nonsystematic generalizations were produced on 85% of test trials. For 2•X and 3•X, nonsystematic generalizations were produced on 13% of test trials each. When the attention parameter was boosted to 1.0, the number of nonsystematic generalizations dropped. For 1•X, nonsystematicities were produced on 50% of test trials. For 2•X and 3•X, nonsystematicities were produced on 8% of test trials each.

The nonsystematic generalizations produced by the network occurred only on generalization test cases. As in the first simulation, these responses corresponded to activating the specific regularity.

The noise and attention parameters can be manipulated to produce different levels of systematicity, but the basic effects of training set (1•X vs. 2•X and 3•X) and low versus high attention remain. The particular systematicity levels do not correspond to any specific experiment, rather, they simply demonstrate the effects of the training set and attention parameters. Detailed simulation of data is left for future work.

Discussion This simulation makes two points. First, the responses due to noise in the input signal were not random, instead they looked like responses to trained stimuli; they were regularizations. This finding is intriguing since the noise is random. Consider the 1•10 case. Normally, the associative constraints from 1• to •1 is suppressed by the identity

constraint from •10 to •10. When the input pattern is perturbed by noise, it becomes more similar to a large number of patterns. The activation of •10 is reduced and the suppression of •1 is also reduced, allowing it to become more activated by 1•. Though this explanation is sensible, much careful work remains to fully understand why random noise produces these nonrandom effects.

The second point is that noisy input signals produced more regularizations when attention is low. The difference was not due to a relative increase in the amount of noise compared to the signal since attention modifies the noise and signal equally. Instead, the explanation lies in the dynamics of the network. Specifically, it lies in the non-linear activation function of the hidden and output units.

When the net input to a unit falls in the middle of the unit's dynamic range, differences are preserved. On the other hand, when the net input to a unit falls near either extreme of that unit's dynamic range, differences between input values are attenuated (See figure 2.). Under attentive, vigilant processing, when the attention parameter is set to 1.0, processing occurs over the middle of each unit's dynamic range, so the signal remains clear despite any background noise. However, when attention is low, processing occurs over the insensitive lower extreme of each unit's range, so the signal becomes murky. A murky signal, apparently, does not produce a strong systematic response, so the default specific regularity response shows up.

General discussion

The phenomenon of systematicity in language and thought has been used to argue for a general purpose, context free, symbol-system cognitive architecture (Fodor & Pylyshyn, 1988). Systematicity, it is argued, *requires* a symbol-system architecture. So how would a symbol-system account for results like those found here? If novel instances are processed systematically, where would the specific regularity ghosts and regularizations come from? These phenomena are not based on superficial perceptual similarity. Rather, they are based on semantic/pragmatic similarity that occurs deep within the cognitive system. To explain on-line comprehension

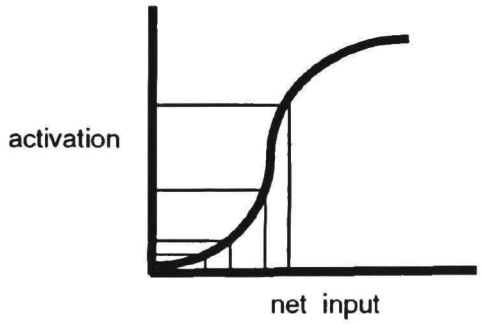

Figure 2. Effect of a non-linear activation function.

regularizations, such as the Moses question, would seem to require part of the input, Moses, to be processed partially and then abandoned or rejected, perhaps on pragmatic grounds but still outside of subjects' awareness. Regularizations in recognition memory require an explanation of why a systematic processor would permit partial matches to memory probes to influence recognition judgements.

But perhaps systematicity is not a fundamental property of the cognitive system. Perhaps instead, it is a more-or-less achievable characteristic of certain large corpora. The idea these simulations support is that the cognitive system is a distributed processor that may achieve *quasi*-systematic performance under appropriate conditions.

Two conditions were explored. First, the corpus must itself be systematic. Each feature of the stimuli must be paired with every other feature so that no context sensitive regularities can be induced. The closer the corpus comes to this criterion, the more systematic will be the network's generalization performance.

When the corpus systematicity requirement is not met, the network will induce specific regularities, associative constraints, between stimulus features. If these regularities are strong, the network will respond to novel stimuli according to these specific regularities. It will produce the responses it learned from trained stimuli. It will regularize. These specific regularities are also valuable for drawing inferences and completing incomplete stimuli with default values.

Larger more systematic corpora produce more systematic generalizations because they strengthen the identity constraints relative to the associative constraints. Yet even in these large corpora, when nonsystematic responses occur, they tend to be the specific regularity responses and they occur in spaces of the corpus where the associative constraints are stronger.

More nonsystematic responses are produced when noise is added to the input signal. Adding background noise to the model is sensible for three reasons: 1) noise is likely to be a quality of the brain, 2) noise provides a mechanism for choosing responses, and 3) noise sets the stage for effects of attention. Interestingly, the nonsystematic responses produced under normally distributed noise conditions are not random but correspond again to specific regularities. Noise, then, lowers the degree of systematicity. A fun phonological example is the "telephone game" where a sentence is *whispered* down a line of people. By the end, the sentence has often changed dramatically.

Finally, modulations of attention can attenuate the effect of noise and return the network's responding to a higher degree of systematicity. In combination with a non-linear activation function, attention can boost a signal nearly obscured by

noise. The view of attention used here is processing power or vigilance to the task.

Several different methods of implementing attention have been used with connectionist models. Cohen, Dunbar, and McClelland (1990) added a constant to the units' net input. This change moves the input to a sharper part of the activation function. Here, attention is a multiplier of the input. Servan-Schreiber, Printz, and Cohen (1990) directly changed the slope of the activation function. Though different in detail, each produces the same essential effect of sharpening the decision threshold of units.

These results come from a very simple and abstract function-learning task. How might these results be extended to more complex cognitive tasks? The extension to language comprehension has been discussed in St. John and McClelland (1990) and St. John (1992, simulation 2). The inputs might be sequences of words and the outputs might be their corresponding events. When the corpus is systematic, novel sentences will be understood systematically. Each word maps straightforwardly onto its associated meaning. However, when specific regularities are present in the corpus, such as semantic regularities, the model may misinterpret a novel sentence to be the closest known meaning. These misinterpretations will occur more frequently when processing is too rapid or scanty. Less well attended information will be processed less well and potentially overridden by specific regularizations from better processed parts of the sentence. Of course, natural language is tremendously complex, so these ideas are simply suggestive. But the phenomena referred to in the introduction indicate that these ideas may be important to understanding language comprehension.

For recognition memory, the corpus corresponds to the training set. During testing, the output corresponds to a familiarity judgement. The closer the output matches the input, the stronger the familiarity score. For novel stimuli, systematic responses will match the input, produce high familiarity scores, and will therefore correspond to false alarms. Research in our lab shows that in a recognition memory task using systematic training corpora, subjects produce a large number of such false alarms.

The results described here are exploratory, but potentially very useful for explaining the quasi-systematicity found in language and memory. Two factors found to affect the degree of systematicity are the degree of systematicity in the training corpus and the amount of attention paid to the task.

References

Brouse, O. and Smolensky, P. (1989). Virtual memories and massive generalization in connectionist combinatorial learning. In *Proceedings of the 11th Annual Meeting of the Cognitive Science Society*: 380-387. Hillsdale, NJ: Erlbaum.

Cohen, J. D., Dunbar, K. & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97,* 332-361.

Erickson, T. D. & Mattson, M. E., (1981). From words to meaning: A semantic illusion, *Journal of Verbal Learning and Verbal Behavior, 20,* 540-551.

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

Goldberg, A. (1992). Argument structure constructions. Dissertation. Berkeley: University of California, Department of Linguistics.

Graesser, A. C. (1981). *Prose comprehension beyond the word.* New York, NY: Springer-Verlag.

Just, M. & Carpenter, P. (1987). *The psychology of reading and language comprehension,* Chapter 14: Speed reading. Newton, MA: Allyn and Bacon.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology: Vol. I.* New York: Wiley.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23,* 1-44.

Medin, D. L. & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review, 85,* 207-238.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39-57.

Pinker, S. (1989). *Learnability and cognition.* Cambridge, MA: MIT Press.

Rumelhart, D. & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & the PDP research group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press.

Seidenberg, M. S. & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523-568.

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science, 249,* 892-895.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317-1323.

St. John, Mark F. (1992). The story gestalt: A model of knowledge intensive processes in text comprehension. *Cognitive Science, 16,* 271-306.

St. John, M. F. & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 217-257.