Strong Systematicity in Sentence Processing by an Echo State Network*

Stefan L. Frank

Nijmegen Institute for Cognition and Information, Radboud University Nijmegen
P.O. Box 9104, 6500 HE Nijmegen, The Netherlands
S.Frank@nici.ru.nl

Abstract. For neural networks to be considered as realistic models of human linguistic behavior, they must be able to display the level of systematicity that is present in language. This paper investigates the systematic capacities of a sentence-processing Echo State Network. The network is trained on sentences in which particular nouns occur only as subjects and others only as objects. It is then tested on novel sentences in which these roles are reversed. Results show that the network displays so-called strong systematicity.

1 Introduction

One of the most noticeable aspects of human language is its systematicity. According to Fodor and Pylyshyn [1], this is the phenomenon that 'the ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others' (p. 37). To give an example, any speaker of English who accepts 'Capybaras eat echidnas' as a grammatical sentence, will also accept 'Echidnas eat capybaras', even without knowing what capybaras and echidnas are.[1]

The ability of neural networks to behave systematically has been fiercely debated [1, 2, 3, 4]. The importance of this discussion to cognitive science is considerable, because neural networks must be able to display systematicity in order to be considered viable models of human cognition. Moreover, it has been argued that, for connectionist systems to *explain* systematicity, they should not just be able to behave systematically but do so as a necessary consequence of their architecture [5, 6].

In a paper investigating sentence processing by neural networks, Hadley [3] defined systematicity in terms of learning and generalization: A network displays systematicity if it is trained on a subset of possible sentences and generalizes to new sentences that are structurally related to the training sentences. The degree

[*] This research was supported by grant 451-04-043 of the Netherlands Organization for Scientific Research (NWO).

[1] Capybaras are large South-American rodents and echidnas are egg-laying mammals that live in Australia.

S. Kollias et al. (Eds.): ICANN 2006, Part I, LNCS 4131, pp. 505–514, 2006.

of systematicity displayed by the network depends on the size of the discrepancy between training and test sentences the network can handle. In particular, Hadley distinguishes *weak* and *strong* systematicity. A network only shows weak systematicity if all words in the test sentences occur in the same 'syntactic positions' that they occupied in the training sentences. Christiansen and Chater [7] correctly note that these 'syntactic positions' are not properly defined, but for the purpose of this paper the term can be taken to refer to a noun's grammatical role (subject or object) in the sentence. Unlike weak systematicity, strong systematicity requires the network to process test sentences with words occurring in *new* syntactic positions. Also, these test sentences must have embedded clauses containing words in new syntactic positions.

An example might clarify this. Suppose a network has only been trained on sentences in which female nouns ('woman', 'girls') occur as subjects and male nouns ('man', 'boys') occur as object. Examples of such sentence are 'woman likes boys' and 'girls see man that woman likes'. The network displays weak systematicity if it can process untrained sentences that (like the training sentences) have female subject(s) and male object(s), such as 'woman likes man' and 'girls that like boys see man'. The network is strongly systematic if it can handle sentences with embedded clauses that have male subject(s) and female object(s), that is, in which the roles of males and females are reversed. Examples of such sentences are 'boys like girls that man sees' and 'man that likes woman sees girls'.

In 1994, Hadley [3] argued that the neural networks of that time showed weak systematicity at best, while strong systematicity is required for processing human language. Since then, there have been several attempts to demonstrate strong systematicity by neural networks, but these demonstrations were either restricted to a few specific test items [7] or required representations [8] or architectures [9] that were tailored specifically for displaying systematicity. In contrast, this paper demonstrates that strong systematicity on a large number of test sentences can be accomplished by using only generally applicable representations and architectures.

Instead of the common Simple Recurrent Network (SRN) [10], an adaptation to Jaeger's [11, 12] Echo State Network (ESN) shall be used because it has been shown to outperform an SRN when weak systematicity is required [13]. Presumably, this is because fewer network connections are trained in an ESN than in an SRN. Since this comes down to a smaller number of parameters to fit, generalization (and thereby systematicity) is improved. It is likely that an ESN will again be superior to an SRN when the task requires strong rather than just weak systematicity.

2 Setup of Simulations

In connectionist cognitive models, sentence processing most often comes down to performing the word-prediction task. In this task, the network processes sentences one word at a time and, after each word, should predict which word will be the next input. The network is successful if it does not predict any word that would form an ungrammatical continuation of the input sequence so far.

The word-prediction task has also been used to investigate the issue of systematicity in neural networks [7, 8, 9, 13, 14], and it shall be used here as well. Section 2.1 describes the language on which the network is trained. After training, the network, presented in Sect. 2.2, is tested on a specific set of new sentences (see Sect. 2.3) to probe its systematic capacities.

2.1 The Language

The network learns to process a language that has a 30-word lexicon, containing 9 singular and 9 plural nouns (N), 5 singular and 5 plural transitive verbs (V), the relative pronoun 'that', and an end-of-sentence marker that is denoted [end] (and also considered a word).

The training sentences are generated by the grammar in Table 1. There is no upper limit to sentence length because of relative clauses that can be recursively embedded in the sentence. Relative clauses, which are phrases beginning with the word 'that', come in two types: subject-relative clauses (SRCs) and object-relative clauses (ORCs). In a SRC, 'that' is followed by a verb phrase (VP), while in an ORC 'that' is followed by a noun. In the training sentences, 20% of the noun phrases (NP) contains a SRC and 20% contains an ORC. This results in an average sentence length of 7 words.

Table 1. Production rules for generating training sentences. Variable n denotes number: singular (sing) or plural (plu). Variable r denotes noun role: subject (subj) or object (obj).

Head		Production
S	→	$\mathrm{S}_{\mathrm{sing}}$ | $\mathrm{S}_{\mathrm{plu}}$
S_n	→	$\mathrm{NP}_{n,\mathrm{subj}}$ VP_n [end]
$\mathrm{NP}_{n,r}$	→	$\mathrm{N}_{n,r}$ | $\mathrm{N}_{n,r}$ SRC_n | $\mathrm{N}_{n,r}$ ORC
VP_n	→	V_n $\mathrm{NP}_{\mathrm{sing,obj}}$ | V_n $\mathrm{NP}_{\mathrm{plu,obj}}$
SRC_n	→	that VP_n
ORC	→	$\mathrm{ORC}_{\mathrm{sing}}$ | $\mathrm{ORC}_{\mathrm{plu}}$
ORC_n	→	that $\mathrm{N}_{n,\mathrm{subj}}$ V_n
$\mathrm{N}_{\mathrm{sing,subj}}$	→	woman | girl | dog | cat | mouse | capybara | echidna
$\mathrm{N}_{\mathrm{plu,subj}}$	→	women | girls | dogs | cats | mice | capybaras | echidnas
$\mathrm{N}_{\mathrm{sing,obj}}$	→	man | boy | dog | cat | mouse | capybara | echidna
$\mathrm{N}_{\mathrm{plu,obj}}$	→	men | boys | dogs | cats | mice | capybaras | echidnas
$\mathrm{V}_{\mathrm{sing}}$	→	likes | sees | swings | loves | avoids
$\mathrm{V}_{\mathrm{plu}}$	→	like | see | swing | love | avoid

As can be seen from Table 1, nouns refer to either females, males, or animals. In training sentences, females ('woman', 'women', 'girl', and 'girls') occur only as subjects while males are always in object position. Animals can occur in either position. In test sentences, the roles of male and female nouns will be reversed (see Sect. 2.3). This means that such sentences *are* considered grammatical, they are just not training sentences, which is why they are not generated by the grammar of Table 1. To define the language *in general*, the four rewrite rules for nouns (N) in Table 1 are replaced by:

$\mathrm{N}_{\mathrm{sing}}$ → woman | girl | man | boy | dog | cat | mouse | capybara | echidna
$\mathrm{N}_{\mathrm{plu}}$ → women | girls | men | boys | dogs | cats | mice | capybaras | echidnas

A note is in place here about the meaning of the terms ‘subject’ and ‘object’, which can sometimes be unclear. Take, for instance, the sentence ‘girls that woman likes see boys’. In the main clause of this sentence, ‘girls’ is the subject (because girls do the seeing) but the same word is object in the subordinate clause (because girls are being liked). To decide upon the noun’s syntactic position in such cases, the method by [7] is used. For the sentences used here this comes down to: Nouns directly following a verb are objects, and all the others are subjects. Table 2 shows some examples of training sentences and indicates which nouns are subjects and which are objects.

Table 2. Examples of training sentences. Subscripts ‘subj’ and ‘obj’ indicate the sentences’ subject(s) and object(s), respectively.

Type	Example sentence
Simple	girls$_{\mathrm{subj}}$ like cat$_{\mathrm{obj}}$ [end] cat$_{\mathrm{subj}}$ likes boy$_{\mathrm{obj}}$ [end]
SRC	girl$_{\mathrm{subj}}$ that likes boys$_{\mathrm{obj}}$ sees cat$_{\mathrm{obj}}$ [end] girls$_{\mathrm{subj}}$ like cats$_{\mathrm{obj}}$ that see boy$_{\mathrm{obj}}$ [end]
ORC	girls$_{\mathrm{subj}}$ that cat$_{\mathrm{subj}}$ likes see boys$_{\mathrm{obj}}$ [end] girls$_{\mathrm{subj}}$ like boy$_{\mathrm{obj}}$ that cat$_{\mathrm{subj}}$ sees [end]
SRC and ORC	girl$_{\mathrm{subj}}$ that likes boys$_{\mathrm{obj}}$ sees cats$_{\mathrm{obj}}$ that man$_{\mathrm{subj}}$ avoids [end] girl$_{\mathrm{subj}}$ that likes boys$_{\mathrm{obj}}$ that cats$_{\mathrm{subj}}$ see avoids man$_{\mathrm{obj}}$ [end]

2.2 The Network

Network Processing. Sentences are processed by an Echo State Network that has been extended with an additional hidden layer, resulting in a total of four layers. These are called the input layer, dynamical reservoir (DR), hidden layer, and output layer, respectively. The input and output layers have 30 units each, corresponding to the 30 words of the language. The DR has 1 000 linear units, and the hidden layer has 10 sigmoidal units. This network is nearly identical to Frank’s [13], who showed empirically that the extra hidden layer and linear DR units are needed for successful generalization in the word-prediction task. The only difference is that the current network has a much larger DR because it needs to process longer sentences and learn long-distance dependencies between nouns and verbs.

Sentence processing by this ESN is similar to that of a four-layer SRN. At each time step (indexed by t), one word is processed. If word i forms the input at time step t, the input activation vector $\boldsymbol{a}_{\text{in}}(t) = (a_{\text{in},1}, \ldots, a_{\text{in},30})'$ has $a_{\text{in},i} = 1$ and $a_{\text{in},j} = 0$ for all $j \neq i$. The output activation after processing the word is computed from the input according to

$$\begin{aligned}
\boldsymbol{a}_{\text{dr}}(t) &= \boldsymbol{W}_{\text{in}}\boldsymbol{a}_{\text{in}}(t) + \boldsymbol{W}_{\text{dr}}\boldsymbol{a}_{\text{dr}}(t-1) \\
\boldsymbol{a}_{\text{hid}}(t) &= \mathbf{f}\left(\boldsymbol{W}_{\text{hid}}\boldsymbol{a}_{\text{dr}}(t) + \boldsymbol{b}_{\text{hid}}\right) \\
\boldsymbol{a}_{\text{out}}(t) &= \mathbf{f}_{\text{sm}}\left(\boldsymbol{W}_{\text{out}}\boldsymbol{a}_{\text{hid}}(t) + \boldsymbol{b}_{\text{out}}\right) \quad ,
\end{aligned}$$

where $\boldsymbol{a}_{\text{in}}, \boldsymbol{a}_{\text{dr}}, \boldsymbol{a}_{\text{hid}}, \boldsymbol{a}_{\text{out}}$ are the activation vectors of the input layer, DR (with $\boldsymbol{a}_{\text{dr}}(0) = \mathbf{0}$), hidden layer, and output layer, respectively; $\boldsymbol{W}$ are the corresponding connection-weight matrices; $\boldsymbol{b}$ are bias vectors; $\mathbf{f}$ is the logistic activation function; and $\mathbf{f}_{\text{sm}}$ is the softmax activation function. As a result of applying $\mathbf{f}_{\text{sm}}$, the total output activation equals 1 and each $a_{\text{out},i}$ can be interpreted as the network's estimated probability that the next input will be word i.

Network Performance. The network's performance is defined as follows: Let G denote the set of words that can grammatically follow the current input sequence, that is, any word $i \notin G$ would be an incorrect prediction at this point. Moreover, let $a(G) = \sum_{i \in G} a_{\text{out},i}$ be the total amount of activation of output units representing words in G, that is, the total 'grammatical' activation.

Ideally, $a(G) = 1$ when there is no 'ungrammatical' output activation. In that case, the performance score equals $+1$. Likewise, performance is -1 if $a(G) = 0$ (there is no grammatical activation). By definition, performance equals 0 if the network learned nothing except the frequencies of words in the training set. If $fr(G)$ denotes the total frequency of the words in G, performance equals 0 if $a(G) = fr(G)$. All in all, this leads to the following definition of performance:

$$\text{performance} = \begin{cases} \frac{a(G) - fr(G)}{1 - fr(G)} & \text{if } a(G) > fr(G) \\ \frac{a(G) - fr(G)}{fr(G)} & \text{otherwise} \ . \end{cases} \tag{1}$$

Network Training. The most important difference between the network used here and an isomorphic SRN is that in ESNs, connection weight matrices $\boldsymbol{W}_{\text{in}}$ and $\boldsymbol{W}_{\text{dr}}$ are not trained but keep their initial random values. All other weights (i.e., those in $\boldsymbol{W}_{\text{hid}}, \boldsymbol{W}_{\text{dr}}, \boldsymbol{b}_{\text{hid}}$, and $\boldsymbol{b}_{\text{out}}$) were trained using the backpropagation algorithm, with a learning rate of .01, cross-entropy as error function, and without momentum. All initial weights and biases were chosen randomly from uniform distributions in the following ranges: $\boldsymbol{W}_{\text{hid}}, \boldsymbol{W}_{\text{out}}, \boldsymbol{b}_{\text{hid}}, \boldsymbol{b}_{\text{out}} \in [-0.1, +0.1]$ and $\boldsymbol{W}_{\text{in}} \in [-1, +1]$. Of the DR connections, 85% was given a zero weight. The other 15% had uniformly distributed random weights such that the spectral radius of $\boldsymbol{W}_{\text{dr}}$ equalled .7.

Ten networks were trained, differing only in their initial connection weight setting. The training sentences, generated at random, were concatenated into one input stream, so the network also had to predict the word following [end], that is, the next sentence's first word. During training, the performance score over a random but fixed set of 100 training sentences was computed after every 1 000 training sentences. As soon as the average performance exceeded .98, the network was considered sufficiently trained.

2.3 Testing for Strong Systematicity

Test Sentences. For a network to display strong systematicity, it needs to correctly process new sentences that have embedded clauses with words occurring in new syntactic positions. Four types of such sentences, listed in Table 3, constituted the test set. Each has one subject- or object-relative clause that modifies either the first or second noun. As such, the test-sentence types are labelled SRC1, SRC2, ORC1, and ORC2.

Table 3. Examples of test sentences of four types

Relative clause		Test sentence	
Type	Position	Type	Example
subject	first	SRC1	boy that likes girls sees woman [end]
	second	SRC2	boy likes girls that see woman [end]
object	first	ORC1	boys that man likes see girl [end]
	second	ORC2	boys like girl that man sees [end]

Test sentences were constructed by taking the structures of the examples in Table 3, and filling the noun and verb positions with all combinations of nouns and verbs, such that:

- Only male nouns appear in subject positions and only female nouns in object positions (note that these roles are reversed relative to the training sentences and that test sentences contain no animal names);
- The resulting sentence is grammatical (i.e., there is number agreement between a noun and the verb(s) it is bound to);
- The two verbs of SRC2, ORC1, and ORC2 sentences differ in number; In SRC1 sentences, where the verbs must have the same number, the first two nouns differ in number;
- The unbound noun (for SRC1 sentences: the third noun) was singular.

This makes a total of 2 (numbers) $\times$ 2^3 (nouns) $\times 5^2$ (verbs) = 400 test sentences of each of the four types. Before processing any of these, the network was given the input [end], putting the DR-units into the right activation state for receiving the test sentence's first word.

Generalization Score. The performance measure defined in (1) assigns a score of 0 to the outputs of a network that has learned nothing but the frequencies of words in the training set. To rate the network's systematicity, a similar measure is used. Instead of using word frequencies as baseline, however, this measure assigns a score of 0 to the outputs of a hypothetical network that has learned the training set to perfection but is not systematic *at all*.

According to Hadley's [3] definition of systematicity, a complete lack of systematicity is the same as the absence of generalization. What can a non-generalizing network do when confronted with a new input sequence? By definition, it cannot use the complete sequence for predicting the next input since this would require generalization. Therefore, the best it can do is to use the most recent part of the test input that was also present in the training sentences. Assume, for instance, that the test input is the ORC1 sentence 'boys that man likes see girl'. After processing '[end] boys', the network is faced with an input sequence it was not trained on because training sentences always begin with a female noun (i.e., 'boys' never follows [end] in the training sentences). All that the network can do is to base its next-word prediction on the last input only, that is, the word 'boys'. In the training sentences, male nouns were followed by [end] in 50% of the cases. Therefore, the non-generalizing network will, by definition, result in an (incorrect) output activation $a_{\text{out,[end]}} = .5$ at this point.

When the next word enters the network, the input sequence is '[end] boys that'. The two-word sequence 'boys that' has appeared in the training sentences (see Table 2). By definition, the output activations of the non-generalizing network at this point are exactly the probabilities that each of the 30 words follows 'boys that' in the training sentences. Likewise, after the word sequence '[end] boys that man', the network basis its predictions on 'man' only because 'that man' never appears in the training sentences. Again, this results in $a_{\text{out,[end]}} = .5$.

The generalization score is computed by an equation identical to (1) except that $fr(G)$ is replaced by the total grammatical output activation of the hypothetical non-generalizing network. This means that positive scores indicate some degree of generalization. If the network scores positively on each word of the four types of test sentences, it displays strong systematicity.

There are several points in the test sentences were generalization is not required for making grammatical predictions. For instance, after '[end] boys that', the next word must be a noun or a plural verb. In the training sentences, 'boys that' is always followed by a plural verb. The non-generalizing network will therefore make the perfectly grammatical prediction that the next word is a plural verb. At such points, even a perfectly systematic network does no better than a non-generalizing one, so the generalization score is not defined. Note that grammatical predictions are always trivial to make at such points, because generalization is not needed. Therefore, nothing is lost by the occasional absence of a generalization score.

3 Results and Conclusion

Generalization scores, averaged over sentences of each type and over the 10 trained networks, are plotted in Fig. 1. The near-perfect performance on the first word of test sentences is a first indication of systematicity by the network. This first word is always a male noun in subject position, which it never occupied in any of the training sentences. Instead, in 50% of the cases, male nouns in training sentences occur in sentence-final position, being followed by [end]. To make the correct prediction that a sentence-initial male noun is *not* followed by [end], the network must have learned that it is the position of the noun that matters and not its particular (male) identity. As is clear from Fig. 1, the network has succeeded in doing this.

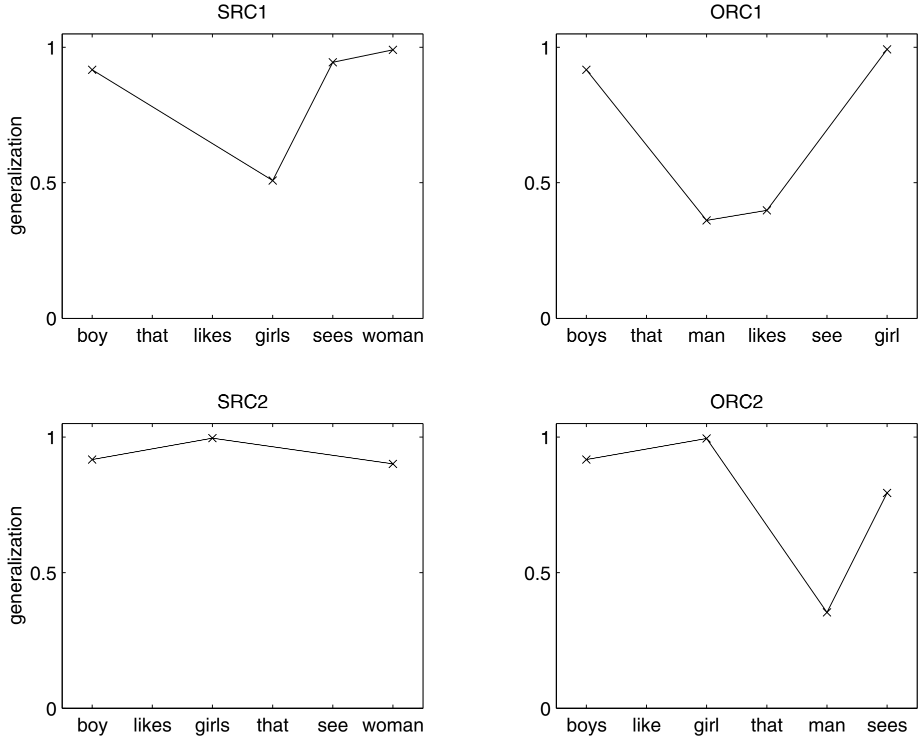

Fig. 1. Generalization scores on test sentences of four types, at all words where generalization is defined (labels on the horizontal axes are only examples of test-sentence words, the plotted performance scores are averaged over all sentences of a type).

For the systematicity to be considered *strong* in the sense of [3], it should also be demonstrated in embedded clauses. The plots in Fig. 1 show that the average performance is above 0 on all words of each test sentence type, including the words in relative clauses. Even at the point in the sentence where performance is minimal, it is still highly significantly positive, as was revealed by sign tests ($p < 10^{-16}$ for each of the four sentences types).

Strictly speaking, these results are not yet proof of systematicity because they could just be an artifact of averaging over networks and sentences. If, for instance, each individual network scores negatively at one point, but this is not the same point for the 10 trained networks, the average generalization score can be positive while none of the networks displays systematicity. The same can happen if each test sentence results in a negative score at some point, but this point differs among the sentences. However, this was clearly not the case: Of all individual generalization scores, only 0.38% was negative.

These results are a clear indication of strong systematicity by the ESN. It processed test sentences with nouns in syntactic positions they did not appear in during training, both in the main clause and in relative clauses, and performed significantly better than a non-systematic network can do. As defined in [3], successful processing of test sentences that differ this much from the training sentences requires strong systematicity.

Nevertheless, generalization decreased considerably at some points in the test sentences. As can be seen from Fig. 1, there are four points at which the generalization score is quite low (less than .6). At all of these problematic points, the network must predict a verb after having processed both a singular and a plural noun. Presumably, the difficulty lies with binding the verb to the correct noun, that is, predicting whether it is singular or plural. An investigation of the network's output vectors supported this interpretation. Nevertheless, the ESN always does better than a network that is not systematic at all and, therefore, displays some (but by no means perfect) strong systematicity.

4 Discussion

Echo State Networks can display not only weak but also strong systematicity. This is a necessary condition for a neural network that is to form a cognitive model of sentence processing. However, Hadley [15, 16] argued that human language is not only systematic with respect to word order in sentences (i.e., syntax) but also with respect to their meaning. That is, people are *semantically systematic*: They can correctly assign a semantic interpretation to sentences they have not been exposed to before. Contrary to this, the word-prediction task used in the simulations presented here is purely syntactic. Recently, however, Frank and Haselager [17] showed that semantic systematicity is not beyond the abilities of a neural network. They trained an ESN to transform sentences into distributed representations of the situations the sentences referred to, and found that the network could generalize to sentences describing novel situations.

References

1. Fodor, J.A., Pylyshyn, Z.W.: Connectionism and cognitive architecture: a critical analysis. Cognition **28** (1988) 3–71
2. Chalmers, D.J.: Connectionism and compositionality: why Fodor and Pylyshyn were wrong. Philosophical Psychology **6**(3) (1993) 305–319
3. Hadley, R.F.: Systematicity in connectionist language learning. Mind & Language **9**(3) (1994) 247–272
4. Niklasson, L.F., Van Gelder, T.: On being systematically connectionist. Mind & Language **9** (1994) 288–302

5. Fodor, J.A., McLaughlin, B.: Connectionism and the problem of systematicity: Why Smolensky's solution does not work. Cognition **35** (1990) 183–204
6. Aizawa, K.: The systematicity arguments. Dordrecht, The Netherlands: Kluwer Academic Publishers (2003)
7. Christiansen, M.H., Chater, N.: Generalization and connectionist language learning. Mind & Language **9**(3) (1994) 273–287
8. Hadley, R.F., Rotaru-Varga, A., Arnold, D.V., Cardei, V.C.: Syntactic systematicity arising from semantic predictions in a Hebbian-competetive network. Connection Science **13**(1) (2001) 73–94
9. Bodén, M.: Generalization by symbolic abstraction in cascaded recurrent networks. Neurocomputing **57** (2004) 87–104
10. Elman, J.L.: Finding structure in time. Cognitive Science **14** (1990) 179–211
11. Jaeger, H.: Adaptive nonlinear system identification with echo state networks. In Becker, S., Thrun, S., Obermayer, K., eds.: Advances in neural information processing systems. Volume 15. Cambridge, MA: MIT Press (2003)
12. Jaeger, H., Haas, H.: Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. Science **304** (2004) 78–80
13. Frank, S.L.: Learn more by training less: systematicity in sentence processing by recurrent networks. Connection Science (in press)
14. Van der Velde, F., Van der Voort van der Kleij, G.T., De Kamps, M.: Lack of combinatorial productivity in language processing with simple recurrent networks. Connection Science **16**(1) (2004) 21–46
15. Hadley, R.F.: Systematicity revisited: reply to Christiansen and Chater and Niklasson and van Gelder. Mind & Language **9**(4) (1994) 431–444
16. Hadley, R.F.: On the proper treatment of semantic systematicity. **14** (2004) 145–172
17. Frank, S.L., Haselager, W.F.G.: Robust semantic systematicity and distributed representations in a connectionist model of sentence comprehension. In Miyake, N., Sun, R., eds.: Proceedings of the 28th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum (in press)