Generalisation towards Combinatorial Productivity in Language Acquisition by Simple Recurrent Networks

Conference Paper · April 2007

DOI: 10.1109/KIMAS.2007.369799

Francis Wong
Nanyang Technological University
10 PUBLICATIONS 92 CITATIONS

Preprint of the paper presented at the 2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems. KIMAS '07: Modeling, Exploration, and Engineering, Waltham, MA, USA.

Generalisation towards Combinatorial Productivity in Language Acquisition by Simple Recurrent Networks

Francis C. K. WONG[a], William S-Y WANG[b]
Language Engineering Laboratory, Department of Electronic Engineering,
The Chinese University of Hong Kong
franciswong@cuhk.edu.hk[a], wsywang@ee.cuhk.edu.hk[b]

Abstract— **Language exhibits combinatorial productivity as complex constructions are composed of simple elements in a linear or hierarchical fashion. Complexity arises as one cannot be exposed to all possible combinations during ontogeny and yet to master a language one need to be, and very often is, able to generalise to process and comprehend constructions that are of novel combinations. Accounting for such an ability is a current challenge being tackled in connectionist research. In this study, we will first demonstrate that connectionist networks do generalise towards combinatorial productivity followed by an investigation of how the networks could achieve that.**

1. INTRODUCTION

Among the theories of child language acquisition one of the central disagreements concerns the nature of the language learning process. On one hand, the nativist school [1, 2] argues that the grammar of language cannot be leant via *deduction* from positive examples alone. Together with the general observation that children receive neither negative examples nor consistent corrective feedbacks from parents, they argue that innate knowledge about language, in the form of abstract rules [3], accounts for the theory of language acquisition. In such a view, the nature of the language learning process is hypothesised as *rule-based* where abstract rules are triggered during ontogeny by exemplars of the target language.

The emergentist school [4-8], however, takes the position that the process of language learning is *statistical-based* in which *induction* from limited examples provides the basis for generalisation to the target language. The evidence supporting this view comes from two sources: the correlation of distributional statistics with syntax and semantics; and experiments of artificial language learning of infants, children, adults and artificial neural networks.

Redington *et al.* [9] extracted distributional statistics, co-occurrence frequencies between lexical items, from corpora of child-directed speech [10] and reported that considerable amount of information about syntactic categories can be obtained from such distributional information alone. Similar approach had been adopted by Li *et al.* [11] on semantics. Various behavioural experiments involving the task of learning/familiarization with artificial languages had been conducted on infants [12, 13], children and adults [14] attempting to demonstrate the statistical nature of the language learning process. However, it remains controversial if the performance of participants in those experiments can also be explained by the learning of rules, as discussed in [15].

2. CONNECTIONIST MODELS OF LANGUAGE PROCESSING

Modelling and simulation work with artificial neural networks (also known as connectionist models), the approach adopted in this study, have been showing success in providing existence proofs of the potential power of statistical-based learning driven by a simple underlying mechanism of association [16-18]. The major merit of the connectionist framework is that the nature of the learning process modelled by the networks is uncontroversially accepted as statistical-based [19-21].

Like many other paradigms in machine learning, the issue of generalisation is also the limiting factor with respect to the success of connectionist models. In the area of the modelling of language acquisition, children exhibit high degree of abilities to generalise. During development they receive only a fraction of the all possible sentences that they will eventually comprehend and produce in adulthood. Criticism [1, 22-24] of the inability of connectionist networks to capture this has often been taken to be the failure of emergentist's statistical-base account of language acquisition.

3. COMBINATORIAL PRODUCTIVITY OF LANGUAGE

One aspect of generalisation that was raised recently by van der Velde *et al.* [25] concerns the ability to deal with the combinatorial productivity of language. Combinatorial productivity is the ability to generalise one's grammatical knowledge about a language in order to comprehend sentences that are composed of novel combinations of lexical items. Take the English simple declarative sentence construction (Noun-Verb-Noun) as an illustration, a sentence is composed of a combination of a subject noun, a main verb and an object noun. The number of such sentences in the language increases with the size of the lexicon raised to the power the length of the sentence. Since the classes of nouns and verbs are open sets and the size of the language grows multiplicatively, the language is consider astronomical in size by van der Velde *et al.* [25].

While humans exhibit the ability to comprehend most, if not all, of the sentences in the language by being exposed to only a fraction of the whole set during ontogeny, successful models of cognition should also exhibit such an ability to generalise combinatorially. Van der Velde *et al.* [25] argued that connectionist models in this regard. Their argument was based on simulation results of simple recurrent networks (SRN, [26]) trained and tested with an artificial language designed to address the issue. However, results that are contrary to van Velde's have been reported independently by us [27] and by Frank [28] in which SRN's potential to deal with combinatorial productivity of language were demonstrated.

Experiments conducted in [25, 27, 28] differed mainly in the architecture of SRNs employed. This leads to our investigation into how such differences influence the performance of SRNs. In this study, on top of arguing that SRNs' potential to generalise had been under-estimated by van der Velde *et al.* [25], we conducted simulations with SRNs of two types of architectures. We have found that SRNs with two hidden layers outperform SRNs with a single hidden layer. We have also probed the question of how networks could achieve generalisation. We hypothesise that networks succeed through the emergence of categories which could be observed by analysing their internal representations developed during the course of training.

4. FRAMEWORK OF ASSESSMENT

A. SRN architectures

Fig. 1 shows the architectures of SRNs used in this study. An SRN without a context layer is just a layered feedforward artificial neural network in which every neuron in a layer is connected via weighted connections to every neuron in the layer immediately above it. When an SRN is to process a sequence of input words, coded in the form of bit vectors, it is fed one word at a time. At the time step t, the hidden layer activation $\boldsymbol{h}(t)$ will be copied via the copy-back connection to the context layer. Such a context layer activation $\boldsymbol{c}$ will then be part of the input to the network in the next time step, i.e. $\boldsymbol{c}(t+1) = \boldsymbol{h}(t)$. In short, the context layer in an SRN provides the network the ability to process sequential information through the accumulation of network's hidden layer activation.

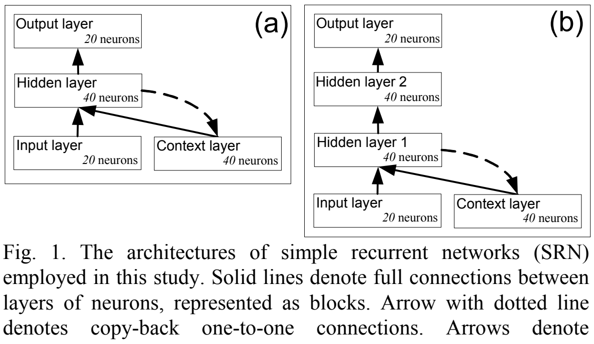

Fig. 1. The architectures of simple recurrent networks (SRN) employed in this study. Solid lines denote full connections between layers of neurons, represented as blocks. Arrow with dotted line denotes copy-back one-to-one connections. Arrows denote directionality. (a) SRN with a single hidden layer; (b) SRN with two hidden layers.

Consistent with [25-28], the two types of SRNs (one with a single hidden layer and the other with two hidden layers) were trained[i] to associate the current word in a sequence with the next word in a sequence given the sentence context in which they appear. In doing so, the network's ability in capturing the grammar of the language can be evaluated by assessing the grammaticality of the network's output in processing a sentence. Take the processing of a simple declarative N-V-N English sentence as an example. Since the network is trained to associate a word in a particular context with the next word, the network's output at the first time step, when a noun is fed, is regarded as the network's *prediction*[ii] of what words to follow. Such a prediction is non-deterministic because virtually all verbs are grammatical continuations of the sentence up to this position. As the lexicon is coded by a set of orthogonal bit strings, the output activation of the SRN is taken to be its estimate of the conditional probability distribution indicating which words to follow. In the literature [25, 29], an error measurement called the Grammatical Prediction Error (GPE) is used to quantify the grammaticality of network's output which is defined as:

$$GPE = 1 - \frac{\sum \text{correct activation}}{\sum \text{correct activation} + \sum \text{incorrect activation}}$$

Continuing with the example of an SRN fed with a partial sentence, the sum of the activations of the output neurons coding for words that are grammatically correct continuations constitutes the numerator in calculating the GPE. The second part of the denominator is obtained in a similar fashion. Notice that the grammaticality of an SRN's prediction requires not just a mere mastery of bi-gram statistics but also the sensitivity to sentence structure. When it comes to the third time step in processing the N-V-N sentence, i.e., when another noun is fed to the network, the network has to take into account the context in which the noun appears in order to differentiate an object noun from a subject noun to achieve a low GPE evaluation at that sentence position.

B. Training and testing data

Following [25], the networks were trained with three types of sentences, simple, right-branching and centre-embedding sentences, as tabulated in Table I. Eight nouns and eight verbs together with the relative marker "that" and the end of sentence marker "#" were incorporated into the lexicon to compose the training and testing sets sentences. The key element behind this framework of assessing SRNs' ability to exhibit combinatorial

TABLE I.
THREE TYPES OF SENTENCE USED IN TRAINING THE SRNS

Sentence types	Constructions	Natural language equivalents
Simple	N-V-N-#	`the boy kisses the girl`
Right-branching	N-V-N-that-V-N-#	`the boy kisses the girl that chases the dog`
Centre-embedding	N-that-N-V-V-N-#	`the girl that the boy kisses chases the dog`

productivity lies in the design of the training and testing sets. We illustrate the rationale with the two *utterance networks* shown in Fig. 2. An utterance, consider only simple sentence construction, is represented by a path through the network from left to right. We consider $\mathcal{L}_C$ as a model of the language available to a child during his acquisition of the target language ($\mathcal{L}_{\mathcal{A}}$). $\mathcal{L}_C$ under-represents the target language in a sense that many of the sentences in $\mathcal{L}_{\mathcal{A}}$ are combinations novel to $\mathcal{L}_C$, e.g n1-v3-n7. For SRNs to be a successful model of language acquisition, it should also exhibit the ability to generalise combinatorially from $\mathcal{L}_C$ to $\mathcal{L}_{\mathcal{A}}$. The training and testing sets sentences were thus constructed accordingly.

The lexicon of nouns and verbs was divided into four non-overlapping groups, we denote the j$^{\text{th}}$ member of the i$^{\text{th}}$ group of nouns as $\mathbf{n}_{\mathbf{ij}}$ and similarly $\mathbf{v}_{\mathbf{ij}}$ for verbs. Training set sentences were composed of nouns and verbs from the same group and hence the complete set of 128 right-branching training sentences are:

Group 1: $\{\mathbf{n}_{1a}\text{-}\mathbf{v}_{1b}\text{-}\mathbf{n}_{1c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{1d}\text{-}\mathbf{n}_{1e}\text{-}\#\}$,
Group 2: $\{\mathbf{n}_{2a}\text{-}\mathbf{v}_{2b}\text{-}\mathbf{n}_{2c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{2d}\text{-}\mathbf{n}_{2e}\text{-}\#\}$,
Group 3: $\{\mathbf{n}_{3a}\text{-}\mathbf{v}_{3b}\text{-}\mathbf{n}_{3c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{3d}\text{-}\mathbf{n}_{3e}\text{-}\#\}$,
Group 4: $\{\mathbf{n}_{4a}\text{-}\mathbf{v}_{4b}\text{-}\mathbf{n}_{4c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{4d}\text{-}\mathbf{n}_{4e}\text{-}\#\}$
where a,b,c,d,e = {1,2}

32 unique sentences (2^5, 2 different words at 5 different syntactic positions) were generated for each group. The other two types of sentences, simple and centre-embedding sentences, were generated in a similar way. The four groups of sentences were combined to form the training sets with different weightings of simple, right-branching and centre-embedding sentences mixed together according to the 4-phased training scheme in Table II. The design of the training scheme with increasing number of complex sentences was in accordance with Elman's notion of "starting small" [30], our pilot simulations have also agreed that training SRNs with simple sentences first, followed by increasing number of complex sentences indeed gives better training results. SRNs trained on training set sentences after the fourth phase of training were evaluated, by GPE metric as introduced in Section 4.A, with testing set sentences.

TABLE II.
4-PHASED TRAINING SCHEME

Phase	Token and type (bracketed) ratio*	No. of sentences fed to a network
1	1 : 0 : 0 (1 : 0 : 0)	32 000
2	6 : 1 : 1 (24 : 1 : 1)	10 240
3	2 : 1 : 1 (8 : 1 : 1)	51 200
4	1 : 2 : 2 (2 : 1 : 1)	640 000

*ratio of simple: right-branching: centre-embedding

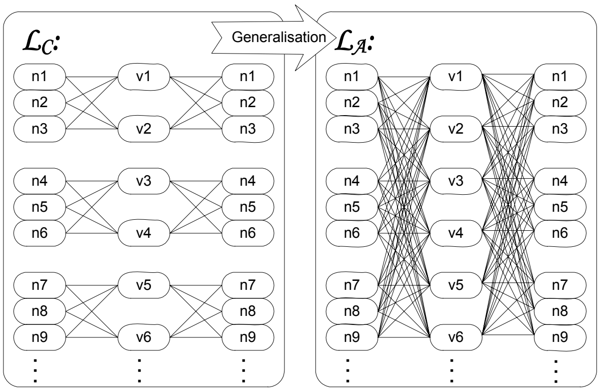

Fig. 2. Combinatorial productivity as generalisation from training set to testing set. Assuming a left-to-right directionality, arrow heads are hidden for simplicity.

The testing sets were constructed by combining lexical items from mixed groups, some of the sentences in $\mathcal{L}_{\mathcal{A}}$ but not in $\mathcal{L}_C$. The level of difficulty with respect to generalisation was varied by the number of groups that were mixed. The more the number of groups that were mixed the more difficult the sentence would be. We use **M** to denote the level of complexity of a testing set sentence. Examples of right-branching testing set sentences with different **M** values are:

M=2: $\{\mathbf{n}_{1a}\text{-}\mathbf{v}_{3b}\text{-}\mathbf{n}_{1c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{3d}\text{-}\mathbf{n}_{1e}\text{-}\#\}$,
$\{\mathbf{n}_{4a}\text{-}\mathbf{v}_{3b}\text{-}\mathbf{n}_{4c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{3d}\text{-}\mathbf{n}_{4e}\text{-}\#\}$
M=3: $\{\mathbf{n}_{1a}\text{-}\mathbf{v}_{3b}\text{-}\mathbf{n}_{2c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{1d}\text{-}\mathbf{n}_{3e}\text{-}\#\}$,
$\{\mathbf{n}_{4a}\text{-}\mathbf{v}_{3b}\text{-}\mathbf{n}_{1c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{4d}\text{-}\mathbf{n}_{3e}\text{-}\#\}$
M=4: $\{\mathbf{n}_{1a}\text{-}\mathbf{v}_{3b}\text{-}\mathbf{n}_{2c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{4d}\text{-}\mathbf{n}_{1e}\text{-}\#\}$,
$\{\mathbf{n}_{4a}\text{-}\mathbf{v}_{3b}\text{-}\mathbf{n}_{1c}\text{-}\mathtt{that}\text{-}\mathbf{v}_{2d}\text{-}\mathbf{n}_{4e}\text{-}\#\}$
where a,b,c,d,e = {1,2}

5. RESULTS ON GPE

Forty SRNs, twenty with a single hidden layer (Fig. 1a) and twenty with two hidden layers (Fig. 1b), were trained. They all attained a very low GPE value with the training set sentences which was consistent with van der Velde *et al.* [25]. The GPE averaged across the twenty networks and across sentence positions attained by one-hidden-layer and two-hidden-layer networks were 0.025 and 0.017 respectively.

Networks' ability to generalise combinatorially was revealed by the GPE evaluation in processing testing set sentences. For each sentence types, 100 testing set sentences of complexity level **M**=3, as simple sentences were of three words in length, were fed to each of the networks. Fig. 3 plots the GPEs at different sentence positions averaged across the twenty networks, as marked on the x-axis, attained by the two types of networks. Results reported by van der Velde *et al.* [25], average GPEs attained by ten networks, are also marked on the plots for comparison. At all sentence positions, we have attained much lower mean GPEs by both architectures

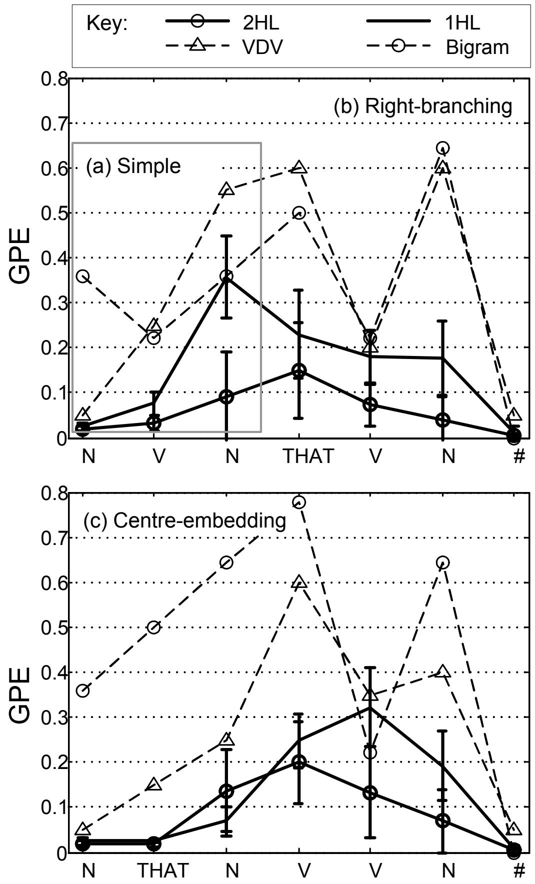

Fig. 3. GPE evaluation on testing set sentences, (a) simple; (b) right-branching and (c) centre-embedding sentences, with complexity level **M**=3. Solid lines with circle markers (labelled 2HL): results from networks with two hidden layers; Solid lines without marker (labelled 1HL): results from networks with a single hidden layer; Dashed lines (labelled Bigram): expected GPE from a bi-gram model; Dashed lines with triangle markers (labelled VDV): results reported by van der Velde *et al.*. Solid lines are results averaged across 20 trials with error bars of two standard deviations in height.

than those of van der Velde's. Networks with two hidden layers showed enhanced performance on generalisation in all but one sentence position. The GPEs expected from a bi-gram model are marked on Fig. 3 as well.

The major criticism raised by van der Velde *et al.* [25] was that SRNs fail to generalise as networks tend to rely on word-to-word association, i.e. bi-gram statistics, in processing novel sentences. They based their argument on the observation that GPEs achieved by their networks were no better than the expected GPEs achieved by a bi-gram model. The results we have obtained clearly are on the contrary. In particular for two-hidden-layer SRNs, in no sentence position the networks achieved a GPE poorer than a bi-gram model. In line with Frank [28], we argue the complaint of van der Velde *et al.* [25] that connectionist models are incapable of dealing with combinatorial productivity is immature.

6. ANALYSIS OF NETWORKS' INTERNAL REPRESENTATIONS

GPE evaluation given in the last section has provided a way to quantify and compare networks' performance on a given task. To look into the possible working mechanism underlying networks' success, analysis of network's internal representations has often been a useful tool.

In an SRN, the hidden layer activation $\boldsymbol{h}(t)$ is a function of: the current word in a sentence (the bit code of a word), context in which the word appear (the context layer activation $\boldsymbol{c}(t)$) and the connection weights to the hidden layer (cf. Section 4.A and Fig. 1a). The mapping from the bit code of a word to the hidden layer activation is equivalent to a mapping of a physical representation of an incoming signal to a functional internal representation in carrying out the prediction task by the network. As the hidden layer activation is of high dimension, 40 in this study as there were 40 hidden layer neurons, methods of dimensionality reduction are employed.

Fig. 4 plots the hidden layer activations of two SRNs, one with a signal hidden layer (plotted in a) and the other with two hidden layers (plotted in b and c). Data points with markers were acquired by feeding the networks with 80 right-branching sentences, 20 of them were training set sentences and 60 testing set sentences of three levels of complexity **M**=2, 3, 4. The group label "N1" in Fig. 4 denotes the hidden layer activations of the networks when the first word of the test sentences, which were nouns, were fed. Similarly "V5" corresponds to the hidden layer activations when the fifth word of the test sentences, which were verbs, were fed. Classical multidimensional scaling (CMS) was used as the method to reduce 40-dimensional data to 2D. We chose CMS instead of principal component analysis (PCA), as employed by some other studies such as [31], because CMS was designed with an objective to match the distance matrix between data points in the reduced space to that in the original high dimensional space [32]. This property could better reveal intrinsic structure, such as categories or clusters, underneath the hidden layer activations. The two networks were chosen for analysis as they attained the lowest mean GPE within their groups of the same type.

It can be observed from Fig. 4 that a more discrete non-overlapping categorisation according to sentence positions was formed in the second hidden layer of a two-hidden-layer network than in the first hidden layer of it or in the sole hidden layer of a one-hidden-layer network. In other words, the classification of main clause subject (N1), main clause object / relative clause subject (N3) and relative clause object (N6) in processing right-branching sentences are more distinctive in the second hidden layer.

Discrete non-overlapping categorisations can also be observed in Fig. 4 (a) and (b) but they tend to be formed according to word types instead. This agrees with the performance difference between the two types of SRNs as grammatical predictions depend not only on the word type of an incoming word but also on the context in which the word appears. Notice that hidden layer activations of V2 and V5 almost completely overlap with one another in Fig. 4 (c). This is due to the current limitation of the prediction task, as in both of these two sentence positions only nouns could be the correct continuation. The task

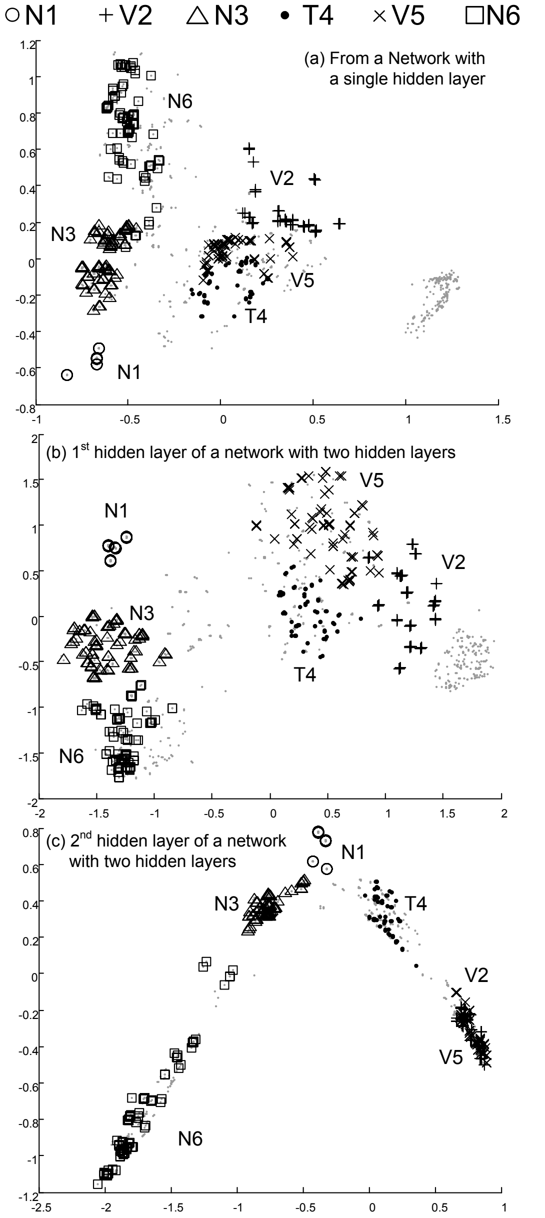

Fig. 4. 2D plots of the hidden layer activations of two networks in the processing of right-branching sentences (N1-V2-N3-T4-V5-N6), (a) the hidden layer activations of a network with a single hidden layer; (b) the 1st hidden layer's activations of a network with two hidden layers; (c) the 2nd hidden layer's activations of a network with two hidden layers. Data points marked with light dots correspond to the hidden layer activations in the processing of other sentence types. The projection from 40-dimensional to 2D was done in classical multidimensional scaling.

does not require the networks to make further distinction.

7. CONCLUSION

In this study, we examined the ability of SRNs to generalise combinatorially to cope with the challenge raised by van der Velde *et al.* [25]. Our simulation results have demonstrated SRNs' ability to achieve high performance with testing set sentences. In particular we have investigated the influence of the network architecture on the performance of the networks. SRNs with two hidden layers outperform those with just a single hidden layer. Analysis of network using CMD was carried out to reveal the potential differential roles played by the two hidden layers, which is a novel contribution of this study. Through the analysis, it has been found that networks that are more successful in generalisation develop internal representations that in one layer are categorised according to general word types and in another layer are categorised according to sentence/syntactic positions. The latter is more specific. We find the failure of one-hidden-layer networks to achieve that most interesting. Does that mean a hierarchical organisation in SRNs is necessary for certain types of generalisation? Is that also true for human cognitive system? It remains to be seen in future research.

ACKNOWLEDGMENT

The authors would like to thank James MINETT, Tao GONG, Susan SHUAI, and Hongying ZHENG for comments and discussions. The research is supported by research grants from the RGC Hong Kong: CUHK-1224/02H and CUHK-1127/04H.

8. REFERENCES

[1] S. Pinker, *Language learnability and language development*, 2nd ed. Cambridge,Mass.: Harvard University Press, 1996.

[2] G. F. Marcus, "Poverty of the stimulus arguments," in *The MIT encyclopedia of the cognitive sciences*, R. A. Wilson and F. C. Keil, Eds. Cambridge, Mass.: MIT Press, 1999.

[3] N. Chomsky, *Aspects of a theory of syntax*. Cambridge, MA: MIT Press, 1965.

[4] B. Macwhinney, "The emergence of linguistic form in time," *Connection Science,* vol. 17, pp. 191-211, 2005.

[5] J. L. Elman, E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett, *Rethinking innateness: a connectionist perspective on development*. Cambridge, Mass.: MIT Press, 1996.

[6] J. L. Elman, "The emergence of language: A conspiracy theory," in *The Emergence of Language*, B. MacWhinney, Ed. Hillsdale, NJ: Lawrence Earlbaum Associates, 1999, pp. 1-27.

[7] M. Tomasello and E. Bates, *Language development: the essential readings*. Malden, Mass.: Blackwell Publishers, 2001.

[8] M. Tomasello, "The item-based nature of children's early syntactic development," *Trends in Cognitive Sciences,* vol. 4, pp. 156-163, 2000.

[9] M. Redington, N. Chater, and S. Finch, "Distributional information: A powerful cue for acquiring syntactic categories," *Cognitive Science,* vol. 22, pp. 425-469, 1998.

[10] B. MacWhinney, *The CHILDES project: tools for analyzing talk*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum, 2000.

[11] P. Li, C. Burgess, and K. Lund, "The acquisition of word meaning through global lexical co-

occurrences," in *Proceedings of the Thirtieth Stanford Child Language Research Forum*, E. V. Clark, Ed. Stanford, CA: Center for the Study of Language and Information, 2000, pp. 167-178.

[12] J. R. Saffran, R. N. Aslin, and E. L. Newport, "Statistical learning by 8-month-old infants," *Science,* vol. 274, pp. 1926-1928, 1996.

[13] L. Gerken, R. Wilson, and W. Lewis, "Infants can use distributional cues to form syntactic categories," *Journal of Child Language,* vol. 32, pp. 249-268, 2005.

[14] J. R. Saffran, "The use of predictive dependencies in language learning," *Journal of Memory and Language,* vol. 44, pp. 493-515, 2001.

[15] M. S. Seidenberg, M. C. MacDonald, and J. R. Saffran, "Does grammar start where statistics stop?," *Science,* vol. 298, pp. 553-554, 2002.

[16] J. L. Elman, "Generalization from sparse input," in *Proceedings of the 38th Annual Meeting of the Chicago Linguistic Society*, 2003.

[17] M. H. Christiansen, C. M. Conway, and S. Curtin, "Multiple-cue integration in language acquisition: a connectionist model of speech segmentation and rule-like behavior," in *Language acquisition, change and emergence: essays in evolutionary linguistics*, J. W. Minett and W. S. Y. Wang, Eds. Hong Kong: City University of Hong Kong Press, 2005, pp. 205-240.

[18] M. H. Christiansen and N. Chater, "Connectionist natural language processing: the state of the art," *Cognitive Science,* vol. 23, pp. 417-437, 1999.

[19] M. Redington and N. Chater, "Connectionist and statistical approaches to language acquisition: a distributional perspective," in *Language acquisition and connectionism*, K. Plunkett, Ed. Hove, UK: Psychology Press, 1998, pp. 129-191.

[20] M. S. Seidenberg and J. L. Elman, "Networks are not 'hidden rules'," *Trends in Cognitive Sciences,* vol. 3, pp. 288-289, 1999.

[21] J. L. McClelland and D. C. Plaut, "Does generalization in infant learning implicate abstract algebra-like rules?," *Trends in Cognitive Sciences,* vol. 3, pp. 166-168, 1999.

[22] S. Pinker and J. Mehler, *Connections and symbols*. Cambridge, Mass.: MIT Press, 1988.

[23] G. F. Marcus, S. Vijayan, S. Bandi Rao, and P. M. Vishton, "Rule learning by seven-month-old Infants," *Science,* vol. 283, pp. 77-80, 1999.

[24] G. F. Marcus and I. Berent, "Are there limits to statistical learning?," *Science,* vol. 300, pp. 53-55, 2003.

[25] F. van der Velde, G. T. van der Voort van der Kleij, and M. de Kamps, "Lack of combinatorial productivity in language processing with simple recurrent networks," *Connection Science,* vol. 16, pp. 21-46, 2004.

[26] J. L. Elman, "Finding structure in time," *Cognitive Science,* vol. 14, pp. 179-211, 1990.

[27] F. C. K. Wong, J. W. Minett, and W. S.-Y. Wang, "Reassessing combinatorial productivity exhibited by simple recurrent networks in language acquisition," in *Proceedings of the 2006 International Joint Conference on Neural Networks* Vancouver, Canada, 2006, pp. 2905-2912.

[28] S. Frank, "Learn more by training less: systematicity in sentence processing by recurrent networks," *Connection Science,* vol. 18, pp. 287-302, 2006.

[29] M. H. Christiansen and N. Chater, "Toward a connectionist model of recursion in human linguistic performance," *Cognitive Science,* vol. 23, pp. 157-205, 1999.

[30] J. L. Elman, "Learning and development in neural networks: the importance of starting small," *Cognition,* vol. 48, pp. 71-99, 1993.

[31] J. L. Elman, "Distributed representations, simple recurrent networks, and grammatical structure," *Machine Learning,* vol. 7, pp. 195-225, 1991.

[32] A. C. Rencher, *Methods of multivariate analysis*: Wiley-Interscience, 2002.

[i] Training in terms of the adjustment of the connection weights was done with the back-propagation algorithm.
[ii] Hence the name “Grammatical Prediction Error” and “prediction task”