

Combinatorial Productivity through the Emergence of Categories in Connectionist 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 — Combinatorial productivity refers to events or objects where complex entities are composed by combining simple elements in a linear or hierarchical fashion. In such cases, complexity in terms of number arises. In human language, sentences are composed of lexicon that may be an ever growing open set. How language learners being exposed to just a fraction of the language generalise their knowledge combinatorially to comprehend all possible grammatical sentences is a challenge being tackled in connectionist modelling research. In this study we will (a) provide simulation experiments to show connectionist networks' potential to generalise combinatorially; (b) explore the plausible mechanism underlining networks' success.

(Keywords: connectionist cognitive sciences, language processing models, recurrent networks, combinatorial productivity)

I. INTRODUCTION

The notion of combinatorial productivity in the present context refers to the abilities to deal with the multiplicative growth of the number of possible combinations of objects with the number of classes of objects to be combined and the sizes of those classes. The idea can be best illustrated with the construction of sentences in a language. Suppose we consider that a sentence is formed by a combination of nouns and verbs in which they will occupy various syntactic positions. The number of possible sentence in the language is m^n , this quantity grows exponentially with the number of syntactic classes, n . In the case of simple declarative S-V-O sentences in English, $n = 3$. The size of the language also grows polynomially with the size of each class, i.e. the number of nouns and the number of verbs in the language, assuming both to be equal to m for simplicity here.

The complexity arisen on one hand poses a challenge in the area of machine learning as often referred "the curse of dimensionality" [3], the problem of finding enough training examples to cope with such a combinatorial complexity with respect to generalization [4]. On the other hand, as recently raised by van der Velde *et al.* [1], it is an issue yet to be properly addressed in building up computational models to account for the combinatorial nature of cognition. The focus of this study is on the latter, we would pay particular attention from a linguistic perspective as we consider that the combinatorial nature of cognition is best expressed in language and more importantly it addresses the old issue of

learnability of language [5, 6] from a new perspective.

A. Combinatorial Productivity in language modelling

For adult language users it is apparently straightforward that having mastered a sentence construction one could comprehend all possible sentences of that type even though most of the instances have near zero probabilities of occurrence in the language input. Adults achieve generalisation from "sparse input" [7] to competence by having mastered the underlying syntactic rules of the language, which governs the lawful and meaningful ways of combining categories of syntactic elements, and the proper assignment of lexical items into syntactic categories. Since rules operate over categories under this traditional framework of analysis, language users by definition possess a high degree of generalisation ability. Once a rule is learnt it immediately applies to every element of the categories that define the rule.

This feature of rules have led some scholars to take rules *per se* as the mechanism underlying language learning [8, 9]. However, in the present study, we would like to seek ways to explain how rules may be deduced from exemplars in the first place. To be more specific, we try to explain the emergence of higher order phenomena from the lower level processing that are more grounded in neurological terms as implemented by artificial neural networks as associations.

We would follow up the discussion raised by van der Velde *et al.* [1, 10, 11] concerning whether connectionist networks as one of the contemporary cognitive models exhibit ability to generalize, to be productive in a combinatorial sense. We would attempt to demonstrate, to the contrary of van der Velde *et al.* [1], that networks do exhibit such abilities and we would also probe the question of how networks could achieve that. 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.

II. CONNECTIONIST LANGUAGE PROCESSING MODELS

The connectionist architecture to be discussed in this study is the simple recurrent network (SRN) model which was proposed by Elman [12] as a model for processing sequential information. It has evolved mainly as models for the acquisition of syntax [13-16] and sentence processing

[17-19].

A. The Network architecture

Fig. 1 shows the general architecture of SRNs that are commonly employed in the literature. Without the context layer, an SRN is just a layered feedforward network in which every neuron in a layer is connected to every other neuron in the layer immediately above it (as shown in Fig. 1, boxed). The context layer in SRNs provides the network the ability to process sequential information through the accumulation of network's hidden layer activation. Suppose

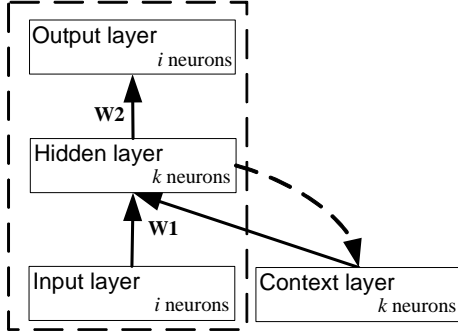


Fig. 1. The general architecture of a simple recurrent network employed in connectionist modelling of language processing. Solid lines denote full connections between layers of neurons, represented as blocks. Arrow with dotted line denotes copy-back one-to-one connections. Arrows denote directionality.

the network is to process a sequence of words ($w_1, w_2 \dots$), the lexicon is by convention encoded with a set of orthogonal bit strings, for example an identity matrix of size i where each column of bit string, $w_{1\dots i}$, codes for a word in the lexicon.

At the first time step, $t=1$, network's input layer activation, $\mathbf{in}(t) \in \mathbb{R}^i$, is set to w_t which is the code of the first word in the sequence. The context layer activation, $\mathbf{c}(t) \in \mathbb{R}^k$, is set to a null context of some neutral value $[0.5, 0.5 \dots 0.5]^T$ since w_t is the first word of the sequence. The concatenation of the vector $\mathbf{in}(t)$ and $\mathbf{c}(t)$ will then be fed to the network and propagated to the hidden layer via the weighted connection $\mathbf{W1}$, which is a $(i+k)$ -by- i matrix. The hidden layer activation, $\mathbf{h}(t) \in \mathbb{R}^k$, is given by the equation, in matrix notation, $\mathbf{h}(t) = \varphi(\mathbf{W1} \times (\mathbf{in}(t) \oplus \mathbf{c}(t)))$, where φ denotes a activation function. In this study the logistic sigmoid function will be used where:

$$\varphi(x) = (1 + e^x)^{-1} \text{ and } \varphi([x_1, x_2 \dots]) = [\varphi(x_1), \varphi(x_2) \dots]$$

Similarly, the output of the network, $\mathbf{o}(t) \in \mathbb{R}^i$, is then obtained by propagating $\mathbf{h}(t)$ to the output layer, i.e. $\mathbf{o}(t) = \varphi(\mathbf{W2} \times \mathbf{h}(t))$. The connection weights, $\mathbf{W1}$ and $\mathbf{W2}$, are modified using the backpropagation algorithm [20, 21] with an objective to minimise the difference between $\mathbf{o}(t)$ and some target output, $\mathbf{p}(t) \in \mathbb{R}^i$, which is associated with $\mathbf{in}(t)$. Very often, a prediction task [12] is used to train the SRNs and hence $\mathbf{p}(t)$ is set to be $w_{(t+1)}$, the next word in the

sequence. This explains why $\mathbf{in}(t)$ and $\mathbf{o}(t)$ are of the same dimension \mathbb{R}^i . More about the prediction task and its use in this study will be discussed in Section II.B and III. Here we focus on giving a brief summary of the working mechanism of SRNs and provide notation for later analysis.

Recall that the context layer in an SRN provides the basis for the network to process sequential data. At the second time step, $t=2$, the context layer activation $\mathbf{c}(t)$ is set to $\mathbf{h}(t-1)$ which is the hidden layer activation of the network at the previous time step. As the process continues though time, the context layer will keep track of the accumulated internal activation of the network. It is commonly denoted as one-to-one copy-back connection between the hidden layer and the context layer, the arrow with dotted line in Fig. 1.

B. Model training and evaluation

As a model for scientific enquiries it should reflect the hypotheses one takes. In the case of SRN model for language processing, it fits into the emergentist school's [15, 16, 22-26] perspectives on plausible language acquisition mechanisms.

First, the usage-based nature of language learning, as proposed in [25, 26], is reflected by the fact that SRNs (and connectionist networks in general) are statistical learning devices and they are trained with positive exemplars alone. Second, minimum assumption is built into the model in order to explore the plausibility for the emergence of linguistic ability out of elementary domain general operations, SRN models attempt to provide an existence proof that syntax can emerge out of temporal associations of sequences of elements. To achieve the latter, the networks are trained to associate the current word in a sequence together with the context in which the word appears with the next word in the sequence, i.e. $\mathbf{in}(t) \oplus \mathbf{c}(t)$ is associated with $w_{(t+1)}$.

A network's ability in capturing the grammar of the language after training can be evaluated by assessing the grammaticality of the network's output in processing a sentence. Take the processing of a simple declarative S-V-O 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*¹ of what words to follow. Such a prediction is non-deterministic because virtually all verbs are grammatical continuations of the partial sentence up to this point. Recall that 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 [1, 27], an error measurement called the Grammatical Prediction Error (GPE) is used to quantify the

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

grammaticality of network’s output which is defined as:

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 predictions 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 S-V-O 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 and to achieve a low GPE evaluation.

In sum, SRNs are faced with the demands of several concurrent and interdependent tasks in learning the artificial language:

- i) forming categories of nouns and verbs, since the lexicon is deliberately coded with orthogonal vectors;
- ii) learning to associate words that are immediately adjacent with each other;
- iii) learning the context dependence of the association.

III. FRAMEWORK OF ASSESSING COMBINATORIAL PRODUCTIVITY

Having introduced the basis of the simple recurrent networks we now turn to the focus of this study, namely, to look at combinatorial productivity exhibited by SRNs. The framework of assessment was introduced by van der Velde *et al.* [1] in which they attempted to demonstrate that SRNs lack the ability to generalise with respect to combinatorial complexity of language and hence argued that SRNs fail to be a model of language acquisition/processing. We have reported our replications of their simulation in [2] showing otherwise. In the remaining parts of this paper we will first summarise our disagreement with van der Velde *et al.* [1, 2] then discuss our recent findings concerning plausible mechanism underlining SRNs’ success in our simulations.

A. Training and testing sets

SRNs of architectures shown in Fig. 3 (a) and Fig. 3 (b) were used in van der Velde [1] and in our previous study [2] respectively. The networks were trained with three types of sentences, simple, right-branching and centre-embedding sentences, as tabulated in Table I. As pointed out by van der

Velde [1], the use of complex sentences would reveal whether networks had truly capture the underlining structure instead of merely bi-gram transitions between nouns and verbs.

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 productivity lies in 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

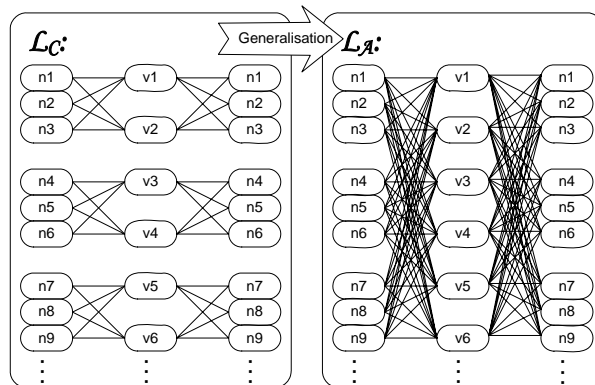


Fig. 2. Combinatorial productivity and generalisation from training set to testing set. Assuming a left-to-right directionality, arrow heads are hidden for simplicity.

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 modelled as \mathcal{L}_A . If we consider an utterance as a combination of lexical items, \mathcal{L}_C under-represents the target language. Such an under-representation parallels the observation of the “skewed input” of the child directed speech in corpus data [28] analysed by Goldberg [29]. Children can and always do generalise their knowledge about the language to comprehend novel sentences like “the horse sings a story”, one of the test stimuli used in [30] as a comprehension task in which children aged 21 to 35 months old showed no problem in understanding in general. For SRNs to be a successful model of language acquisition, it should also exhibit the ability to generalise from \mathcal{L}_C to \mathcal{L}_A . The training and testing sets sentences were thus constructed according to Fig. 2.

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

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

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

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)	64 000

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

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” [1, 31], our initial 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, via GPE as introduced in Section II.B, with testing set sentences.

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

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

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

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

where a.b.c.d.e = {1,2}

Obviously testing sets contain more sentences than the training set. More importantly, constructing testing set sentences this way ensures a maximum separation between lexical items from the same group since GPE is evaluated on every sentence position.

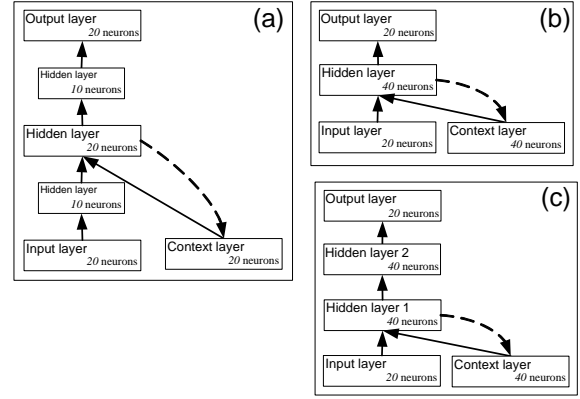


Fig. 3. Network architecture used in (a) van der Velde [1], (b) Wong [2] and (c) this study.

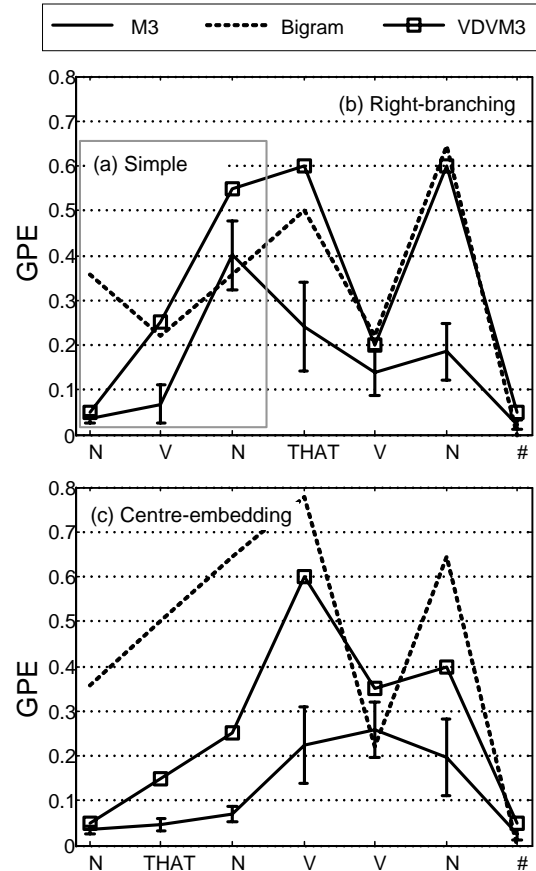


Fig. 4. GPE evaluation on testing set sentences, with complexity level $M=3$. Solid line (labelled M3): results we obtained; Solid line with square marker (labelled VDVM3): results reported by van der Velde *et al.* [1]; Dotted line: expected GPE from a bi-gram model.

IV. OUR RESULTS CONTRARY TO VAN DER VELDE

We trained twenty SRNs with the architecture shown in Fig. 3(b), each initialized with an independent random initial set of connection weights, with streams of concatenated sentences that were randomly sampled from the training

sets. Similar to the results reported by van der Velde *et al.* [1], the networks achieved a small GPE, on average about 0.05, with training set sentences. Since the prime focus is to examine the generalisation ability exhibited by the networks. We plot only testing set GPE in Fig. 4 as a comparison of the results obtained by us [2] and by van der Velde *et al.* [1].

TABLE III.
THE BI-GRAM MODEL*

	N	V	that	#
N	0	0.357	0.286	0.357
V	0.778	0.222	0	0
that	0.5	0.5	0	0
#	1	0	0	0

* the value of the cell in the i^{th} row j^{th} column is the relative frequency that the words in category j follow the word in category i . Formally, $\Pr(w_{k+1} \in C_j | w_k \in C_i)$, where w_k and w_{k+1} are consecutive words in a sequence. They are calculated from the training data.

GPE of networks' output in processing testing set sentences with complexity level $M=3$ in each sentence position were measured. They were averaged over the twenty SRNs and are plotted in Fig. 4 with error bars of two standard deviations in height. Results reported by van der Velde *et al.* [1] are marked on the plots with square markers. A baseline GPE pattern expected from a bi-gram model (Table III) is also included.

Van der Velde *et al.* [1] argued that SRNs are only able to make use of the bi-gram statistics when processing testing set sentences and therefore failed to generalise with respect to combinatorial complexity (consider at the position of the second noun of a right-branching sentence, the frequent N-V transitions would bias the network to predict another verb to follow, which is grammatically incorrect). Their argument was based on the observation that the testing set GPE they obtained were, in most sentence positions, larger than the expected values obtained from a bi-gram model. In other words, they were attempting to show that SRNs fail to capture the sentence structure when the sentence involved novel combinations of lexical items.

Our simulation results, however, showed contrary observation. In most sentence positions, our networks achieved GPE smaller than the bi-gram GPE. This cast doubt on their criticism on SRNs. We speculate that our improvement in training the networks lies mainly in the choice of the network architecture (cf. Fig. 3). SRNs in van der Velde *et al.* [1] had three hidden layers with the recurrent copy-back connections coming from the second hidden layer. In our initial simulations, SRNs with recurrent connections come from layer other than the first hidden layer showed poor performance on testing set sentences, just like what van der Velde *et al.* [1] reported, even though they learned the training set well. This was the reason we adopt the architecture we used. In short, we consider the dismissal of SRNs as a model for language acquisition and processing based on a premature investigation as in van der Velde *et al.* [1] unconvincing. And that triggered us to further explore the impact of network architecture on the performance with

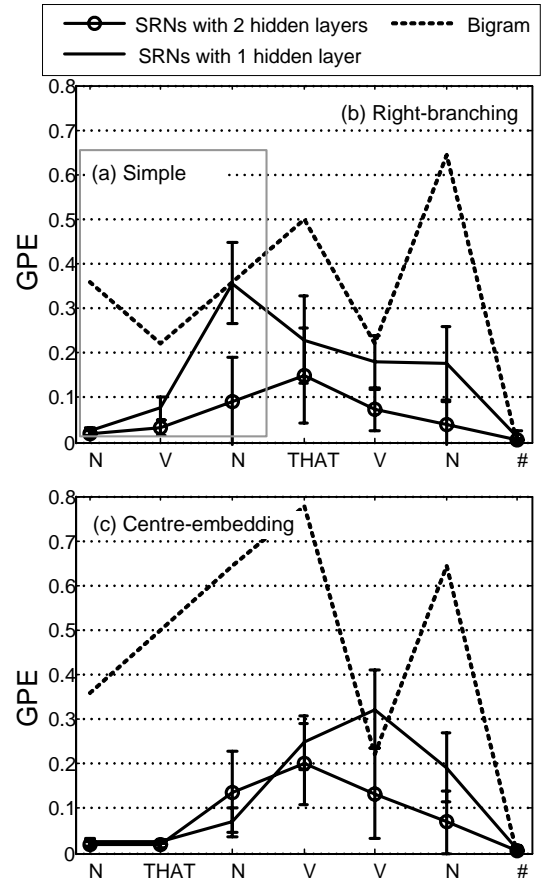


Fig. 5. Comparing the performance on generalisation exhibited by SRNs with two hidden layers and SRNs with one hidden layer.

respect to generalisation.

We carried out another set of simulations with SRN of architecture shown in Fig. 3 (c). Because the number of connection weights for each network was increased by 50 per cent, we extended the fourth phase (cf. Table II) of training such that the number of training sentences fed to the network in the last phase was increased to 640,000. The tenfold increment was to exclude the possibility of premature training that might happen. Twenty 2-hidden-layer-SRNs and twenty 1-hidden-layer-SRNs were trained according to the revised training scheme. GPE evaluation after training, averaged over the twenty trials of each network type, are plotted in Fig. 5. Networks with two hidden layers showed considerable smaller GPE in processing testing set sentences. Although centre-embedding sentences remained to be difficult to process which is consistent with human performance [32, 33]. Networks with only one hidden layer showed no improvement in response to the extended training. The difference in performance between one-hidden-layer networks and two-hidden-layer network together with the individual difference in achieved GPE among networks of the same type lead us to explore if there exist a qualitative difference between the two types of networks that may shed light on the plausible mechanism

underlines generalisation. This is to be discussed in the next section.

V. COMBINATORIAL PRODUCTIVITY THROUGH CATEGORIES

Proponents of connectionist models for language acquisition [7, 34] have long been arguing that networks are more than passive statistics gathering device. Attempts have been made to show that networks could go beyond surface similarity towards successful generalisation. Elman’s early attempt in [12], in which architecture of Fig. 3(b) was used, showed how knowledge of categories of nouns and verbs; sub-categorisation of nouns into animates and inanimates; and sub-categorisation of verbs into transitive and intransitive verbs could be induced by SRNs. The active role played by, and modelled by, neural networks was clearly stated by McClelland and Plaut in [34]:

"The relevant overlap of representations required for generalization in a neural network or other statistical learning procedure need not be present directly in the ‘raw input’ but can arise over internal representations that are subject to learning."

In the context of the current study, a testing set sentence such as “ $\mathbf{n}_{11}\text{-}\mathbf{v}_{21}\text{-}\mathbf{n}_{31}$ ” is novel to the network since the \mathbf{n}_{31} noun was never seen by the network as an object in a sentence with \mathbf{n}_{11} as the subject. Results in Fig. 5 show that SRNs with two hidden layers are better in dealing with such novelty and some of them even achieved a very low GPE value, Fig. 7 plots the mean GPE achieved by each of the networks with two hidden layers. We speculate that on top of forming the categories of nouns and verbs, as demonstrated in Elman [12], categorisation according to sentence positions may also be the driving force for the success of the networks.

In the literature of connectionist research, to probe the question of how the networks solve a task, quite often analysis will be done on the internal representations, the hidden layer activations (denoted as $\mathbf{h}(t)$ in Section II.A), developed by the networks through training. Since hidden layer activations are of high dimension, method of dimensionality reduction such as Principal Components Analysis, Multidimensional Scaling, and Hierarchical Clustering Analysis are often used. We choose to use the Classical Multidimensional Scaling as it target at preserving the Euclidean distance between data points in the reduced space and hence might be better in revealing categories in the form of clusters formed by the networks.

A. Analysis of hidden layer activations – network #8

Fig. 6 gives the scattering plot of hidden layer activations sampled from network #8, the network that achieved the lowest mean GPE (of 0.003, cf. Fig. 7) on $M=3$ testing set sentences among the twenty 2-hidden-layer SRNs trained, i.e. the one that was most able to generalise. The network was fed with 90 simple sentences, 30 for each complexity level, $M=1\dots3$; 120 right-branching and 120 centre-

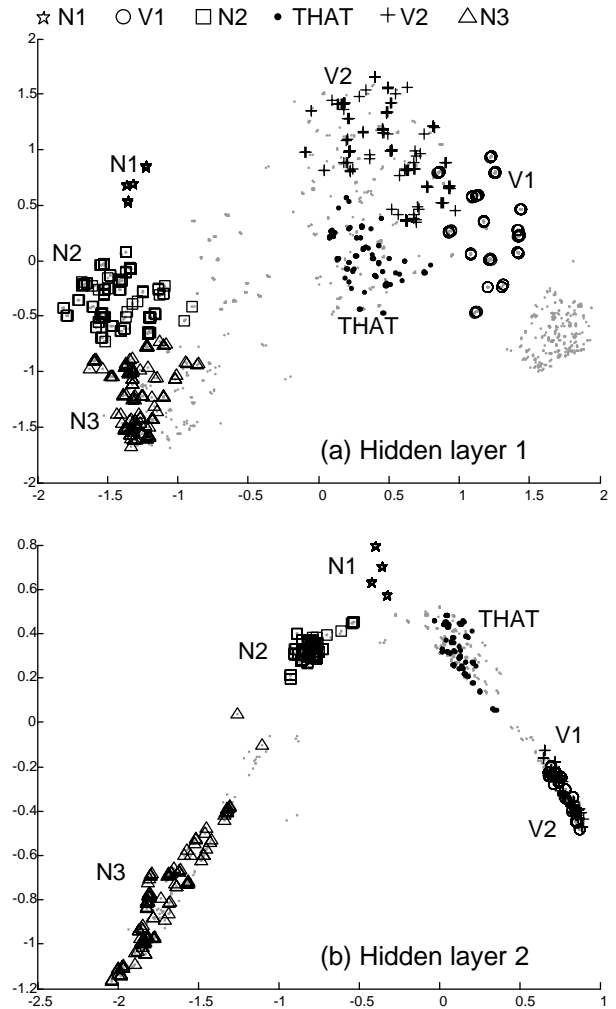


Fig. 6. Hidden layer activations of network #8’s (a) first hidden layer and (b) second hidden layer. Each data point corresponds to a 2D projection of network’s hidden layer activation, i.e. $\mathbf{h}(t)$, in processing a word in a sentence. Data points correspond to activation at different sentence positions of right-branching sentences, “N1-V1-THAT-V2-N2”, are marked with different shape in black and labelled accordingly. Dots in grey correspond to the processing of other two sentence types. The projection was done with Classical Multidimensional Scaling.

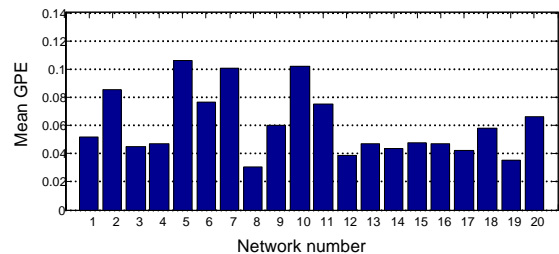


Fig. 7. Mean GPE achieved by each of the 2-hidden-layer networks. Obtained by testing the networks with 300 $M=3$ testing set sentences, 100 for each sentence type.

embedding sentences, 30 for each complexity level, $M=1\dots4$. The network was fed word-by-word with the test sentences and the hidden layer activation $h(t)$ at each time step, which corresponds to each sentence position, was recorded. Since the first three words of a right-branching sentence is equivalent to a simple sentence and due to the limitation of space we highlight only data points that correspond to the processing of right-branching sentences in Fig. 6. The first nouns (N1) in the right-branching sentences are marked with a star, the second noun (N2) are marked with a square and the last nouns (N3) are marked with a triangle in Fig. 6. Labels for the two verb positions and the relative marker ‘that’ position can be found in the figure.

A qualitative difference between the functioning of the first and the second hidden layer can be observed. In the second hidden layer, a more distinct clustering according to sentence positions is observed compared with the groupings formed in the first hidden layer which are more dominated by word type. In other words, the classification of main clause subject (N1), main clause object / relative clause subject (N2) and relative clause object (N3) in processing right-branching sentences are more distinctive in the second hidden layer. This agrees with the low GPE evaluation obtained by the network as grammatical predictions depend not only on the word type of the incoming word but also on the context in which the word appears. Notice that hidden layer activations of V1 and V2 almost completely overlap with one another. 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 does not require the network to make further distinction.

B. Analysis of hidden layer activations – network #5

The worst network in terms of generalisation is network #5 (cf. Fig. 7) which achieved a mean GPE of 0.1 in processing $M=3$ testing set sentences. Fig. 8 shows plots of the network’s hidden layer activations of the two layers. While a somewhat fuzzy distinction between nouns and verbs is still available in the first hidden layer, clear distinction according to sentence positions is lost in the second layer. Analysis of other networks with relatively poor performance, e.g. network #7 and #10, show similar phenomenon.

VI. SUMMARY AND CONCLUSION

We consider the dismissal by van der Velde *et al.* [1, 10, 11] of SRNs as a model for cognition based on a premature analysis of the networks’ performance unconvincing. Our experimentations with SRN, under the framework of combinatorial productivity, suggested that (i) networks do exhibit ability to generalise, (ii) networks with recurrent connections coming from the first hidden layer generalise better than networks with recurrent connection coming from other layer and (iii) networks with two hidden layers show better ability to generalise.

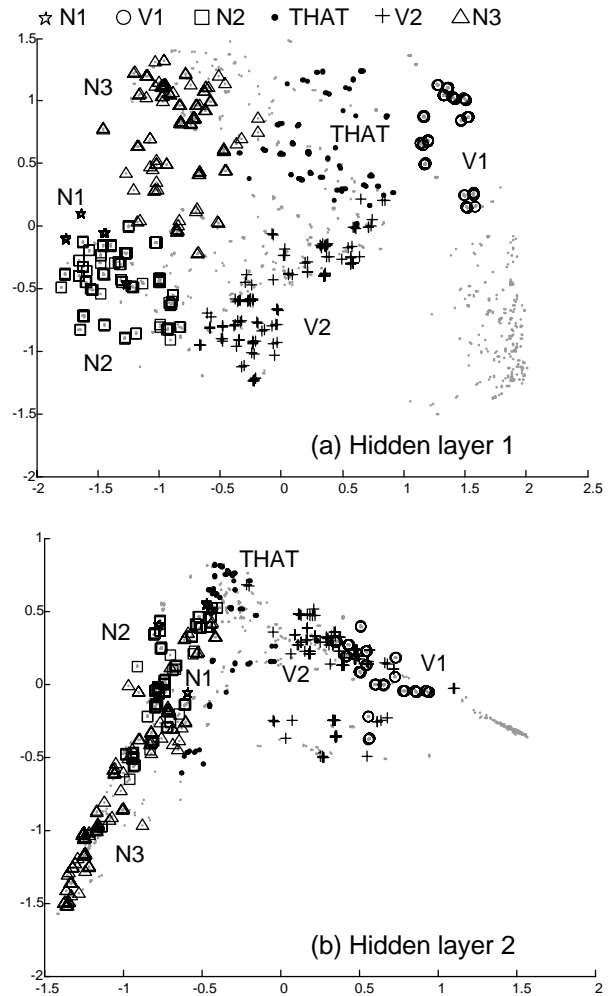


Fig. 8. Hidden layer activations of network #5’s (a) first hidden layer and (b) second hidden layer. See caption of Fig. 6 for details.

To our knowledge, this study is the first one to show that layers could play differential roles in connectionist networks for language modelling. Among the two-hidden-layer networks we analysed, all of them showed various degree of success in forming general nouns and verbs categories in the first hidden layer. SRNs that were more successful with respect to generalisation not only developed a more separated distinction between nouns and verbs in the first hidden layer. They also made fine categorisations according to the sentence context in the second hidden layer. Our speculation that the success of two-hidden-layer SRNs is driven by the categorisation on top of general noun-verb distinction was supported by such observation, particularly from the contrast between the more successful networks with the less successful ones.

From the perspective of psycholinguists, one might ask whether the model and results we presented support a view that noun-verb distinction is first acquired by a child before acquiring the ability to comprehend a sentence. The answer is no as we have not traced the development of categories through time to see what type of categorisation evolved first.

In other words, the analysis done on the two hidden layers was strictly synchronic. Our speculation is that they are likely to be coevolving with one another since sentence context certainly provides a strong cue towards identification of word class. This view is also supported by multi-agent model of language emergence [35, 36].

It remains for future research to see how much of the observed “working mechanism” in SRNs parallels the function of neural substrates in the brain of a language learner. Nevertheless, the functional building block of artificial neural networks is the association between functionally correlated signals. If complex behaviour can emerge out of such low level association in artificial networks, we see no reason why that cannot happen in the living brain.

ACKNOWLEDGEMENT

The authors would like to express their gratitude towards Dr. James W. MINETT and Mr. GONG Tao for useful comments. The research is supported by research grants from the RGC Hong Kong: CUHK-1224/02H and CUHK-1127/04H.

REFERENCES

- [1] 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.
- [2] F. C. K. Wong, J. W. Minett, and W. S.-Y. Wang, "Reassessing combinatorial productivity exhibited by simple recurrent networks in language acquisition," in *2006 International Joint Conference on Neural Networks*, Vancouver, Canada, 2006, pp. 2905-2912.
- [3] R. E. Bellman, *Adaptive control processes*: Princeton University Press, 1961.
- [4] L. I. Perlovsky, "Toward physics of the mind: concepts, emotions, consciousness and symbols," *Physics of Life Reviews*, vol. 3, pp. 23-55, 2006.
- [5] S. Pinker, *Language learnability and language development*, 2nd ed. Cambridge, Mass.: Harvard University Press, 1996.
- [6] G. Lupyán and M. H. Christiansen, "Case, word order, and language learnability: Insights from connectionist modeling," in *Proceedings of the 24th Annual Conference of the Cognitive Science Society* Mahwah, NJ: Lawrence Erlbaum Associates, 2002, pp. 569-601.
- [7] J. L. Elman, "Generalization from sparse input," in *Proceedings of the 38th Annual Meeting of the Chicago Linguistic Society*, 2003.
- [8] G. F. Marcus, *The algebraic mind: integrating connectionism and cognitive science*. Cambridge, Mass. ; London: MIT Press, 2001.
- [9] 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.
- [10] F. van der Velde and M. de Kamps, "Neural blackboard architectures of combinatorial structures in cognition," *Behavioral and Brain Sciences*, vol. 29, pp. 37-108, 2006.
- [11] F. van der Velde, "Modelling language development and evolution with the benefit of hindsight," *Connection Science*, vol. 17, pp. 361-379, 2005.
- [12] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, pp. 179-211, 1990.
- [13] A. Borovsky and J. L. Elman, "Language input and semantic categories: a relation between cognition and early word learning," *Journal of Child Language*, vol. 33, pp. 759-790, 2006.
- [14] 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.
- [15] J. L. Elman, "Connectionism and language acquisition," in *Language development: the essential readings*, M. Tomasello and E. Bates, Eds. Malden, Mass.: Blackwell Publishers, 2001.
- [16] J. L. Elman, "An alternative view of the mental lexicon," *Trends in Cognitive Sciences*, vol. 8, pp. 301-306, 2004.
- [17] M. H. Christiansen and N. Chater, "Connectionist natural language processing: the state of the art," *Cognitive Science*, vol. 23, pp. 417-437, 1999.
- [18] M. H. Christiansen and J. T. Devlin, "Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations," in *Proceedings of the 19th Annual Cognitive Science Society Conference* Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 113-118.
- [19] P. Rodriguez, "Simple recurrent networks learn context-free and context-sensitive languages by counting," *Neural Computation*, vol. 13, pp. 2093-2118, 2001.
- [20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representation," in *Parallel distributed processing: explorations in the microstructure of cognition*. vol. 1, D. E. Rumelhart, J. L. McClelland, and University of California San Diego. PDP Research Group., Eds. Cambridge, Mass.: MIT Press, 1986, pp. 319-362.
- [21] S. S. Haykin, *Neural networks: a comprehensive foundation*, 2nd ed. New York: Prentice Hall, 1999.
- [22] J. L. Elman, "The emergence of language: A conspiracy theory," in *The Emergence of Language*, B. MacWhinney, Ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1999, pp. 1-27.
- [23] 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.
- [24] M. Redington, N. Chater, and S. Finch, "Distributional information: A powerful cue for acquiring syntactic categories," *Cognitive Science*, vol. 22, pp. 425-469, 1998.
- [25] M. Tomasello, "The item-based nature of children's early syntactic development," in *Language development: the essential readings*, M. Tomasello and E. Bates, Eds. Malden, Mass.: Blackwell Publishers, 2001, pp. 169-186.
- [26] M. Tomasello, *Constructing a language: a usage-based theory of language acquisition*. Cambridge, Mass.: Harvard University Press, 2003.
- [27] M. H. Christiansen and N. Chater, "Toward a connectionist model of recursion in human linguistic performance," *Cognitive Science*, vol. 23, pp. 157-205, 1999.
- [28] E. Bates, I. Bretherton, and L. Snyder, *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge, MA: Cambridge University Press, 1988.
- [29] A. E. Goldberg, *Constructions at work: The nature of generalization in language*. New York: Oxford University Press, 2006.
- [30] V. Valian, S. Prasada, and J. Scarpa, "Direct object predictability: effects on young children's imitation of sentences," *Journal of Child Language*, vol. 33, pp. 247-269, 2006.
- [31] J. L. Elman, "Learning and development in neural networks: the importance of starting small," *Cognition*, vol. 48, pp. 71-99, 1993.
- [32] F. Hsiao and E. Gibson, "Processing relative clauses in Chinese," *Cognition*, vol. 90, pp. 3-27, 2003.
- [33] E. Gibson, "The dependency locality theory: a distance-based theory of linguistic complexity," in *Image, language, brain: papers from the First Mind Articulation Project Symposium*, A. Marantz, Y. Miyashita, and W. O'Neil, Eds. Cambridge, Mass.: MIT Press, 2000, pp. 95-126.
- [34] 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.
- [35] T. Gong and W. S. Y. Wang, "Computational modeling on language emergence: A coevolution model of lexicon, syntax and social structure," *Language and Linguistics*, vol. 6, pp. 1-41, 2005.
- [36] T. Gong, J. Ke, J. W. Minett, J. H. Holland, and W. S. Y. Wang, "A computational model of the coevolution of lexicon and syntax," *Complexity*, vol. 10, pp. 50-62, 2005.

¹ Hence the name “prediction task” and “Grammatical Prediction Error”