# Language Processing in Real and Artificial Neural Networks

**WONG, Chun Kit**

A Thesis Submitted in Partial Fulfilment of

the Requirements for

the Degree of Doctor of Philosophy

in Electronic Engineering

March 2009





ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

# Abstract

The first part of this thesis reports a computational simulation work on *modelling language processing*. The motivation was from a study that attempted to show that contemporary connectionist models, the simple recurrent networks being the exemplar, fail to exhibit a kind of generalisation that is essential for acquiring natural language. A replication of the simulation, however, revealed a pattern that was to the contrary. I extended the simulation to investigate the plausible causes of the discrepancies and I argued that generalisation is possible in connectionist models as long as the knowledge of categories is successfully induced by the networks.

The second part of the thesis reports an investigation of how *the real brain functions* during tasks of language processing. To be more specific, during a task of reading for comprehension. I employed a high temporal resolution brain imaging technique known as electroencephalography (EEG), through which the electrical potentials, in terms of micro-volts, as a consequence of participants' engagement of the reading task were measured from electrodes that were placed on their scalps. Event-related-potentials (ERPs), also referred to as "brain waves", were then derived from these measured time

series of voltages. In this ERP study of reading, I identified the earliest brain responses to semantic processing and I argued that within one fifth of a second the brain has already started to process the meanings of a word.

Despite the fact that the two parts of the thesis represent research work of two different disciplines, the central theme of my thesis remains clear—to study how language is possible in a brain.

# 序

本論文首先討論利用電腦模擬之方法去研究語言處理的問題。這部份研究的動機來自一項其他學者之文章。該文指出現今流行的一種連結論模型(connectionist models) — simple recurrent networks，缺乏一種對語言習得很重要的特質，概括能力。本論文先把該文章的實驗重複，發現結果跟原文有所不同。本論文將匯報由此而引申的一系列仿真實驗及其分析。

本論文的第二部份將報告一項腦電圖研究，此研究的目的為探討人在閱讀時大腦是如何運作的。我們把電極放在參加者之頭皮上以量度他們因閱讀而產生的電位反應。此腦電圖研究發現在不到五份一秒的時間，大腦已對文字的語意有所反應。

此兩項研究雖代表著兩門學科的研究方法，但兩者皆指向同一方向 — 探討語言及大腦之間的關係。殊途而同歸。

# Contents

# List of Figures

# List of Tables

# Notations

Vector $\boldsymbol{v} = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}$

Matrix $\mathbf{W} = \begin{bmatrix} w_{11} & \dots & w_{1j} & \dots & w_{1n} \\ \vdots & & & & \vdots \\ w_{i1} & \dots & w_{ij} & \dots & w_{in} \\ \vdots & & & & \vdots \\ w_{m1} & \dots & w_{mj} & \dots & w_{mn} \end{bmatrix}$

# Chapter 1

# Introduction

This thesis reports two lines of research work I have conducted during my days in The Chinese University of Hong Kong.

## 1.1 Chapters overview

**The first part** of the thesis reports my work on *computational modelling of language processing*. The motivation came from a study[1] in which the authors attempted to show that contemporary connectionist models, the simple recurrent networks being the exemplar, fail to exhibit the kind of generalisation that is essential for acquiring language. I started out by replicating the simulation they reported and my simulation results were contrary to the original study. This will be reported in Chapter 2. I then extended the simulation to investigate the plausible causes of the discrepancies, which

[1] van der Velde & de Kamps (2006); van der Velde, van der Voort van der Kleij & de Kamps (2004)

will be discussed in Chapter 3. In Chapter 4, I will demonstrate that generalisation is possible in connectionist models as long as the knowledge of categories is successfully induced by the networks.

**The second part** of the thesis reports an investigation of how *the real brain functions* during tasks of language processing. To be more specific, during a task of reading for comprehension. I employed a brain imaging technique known as electroencephalography (EEG)[2] through which the electrical potentials, in terms of micro-volts, as a consequence of participants' engagement of a task were measured from electrodes that were placed on their scalps.[3] Event-related-potentials (ERPs), also referred to as "brain waves", were then derived from these measured time series of voltages. Chapter 5 will report the ERP study I conducted on Chinese reading in which I attempted to identified the earliest brain responses to semantic processing.

Despite the fact that the two parts of the thesis represent research work of two different disciplines, the central theme of my thesis remains clear—to study how language is possible in a brain. The use of artificial neural networks to model the brain and the use of electroencephalography to observe a brain in action have a striking similarity.

[2]Electro-en-cephalo —electricity-in-the-brain

[3]Figure 1.1 illustrates the use of EEG and the setting I used to conduct my ERP study.

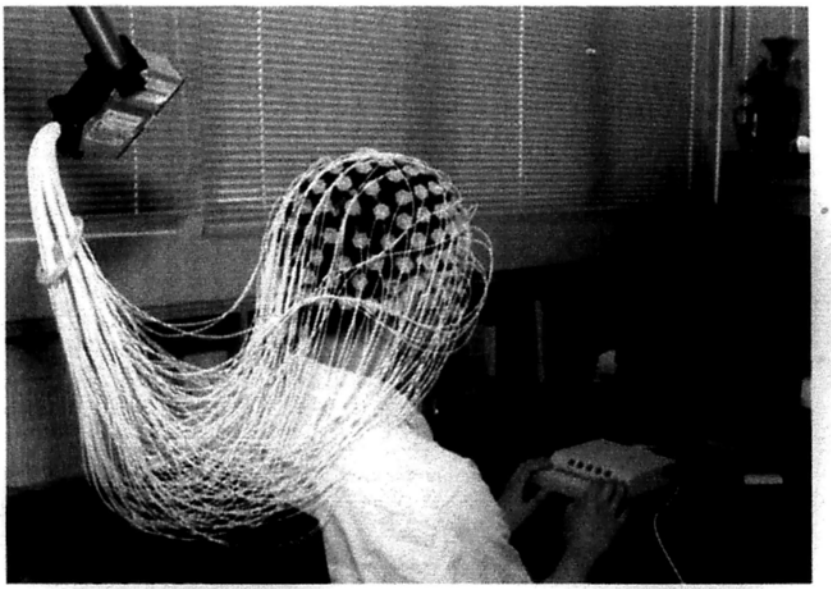


Figure 1.1: The use of EEG to study how the brain works

Connectionist modelling was rooted in Hebb's (1949) "*neurophysiological postulate*" of learning and behaviour in which he put:

> *"When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."*
>
> Hebb (1949, p. 62)

as a hypothesis of the neurological bases of learning. This is later known as the *Hebbian learning* which has led to the development of Perceptrons (Minsky & Papert, 1969) and the backpropagation networks (Rumelhart & McClelland, 1986; McClelland & Rumelhart, 1986) with the very idea that complex behaviour can emerge from the interactions between simple processing units, neurons, that by themselves merely associate local and incomplete information.

On the other hand, electroencephalography measures signals that are hypothesised to be the summed activities of cortical pyramidal neurons in response to the engagement of particular cognitive tasks (Allison, 1984). These brain waves would not have been measurable unless tens of thousands of pyramidal neurons are coherently oriented and synchronously activated (Pizzagalli, 2007).

The use of EEG as a brain imaging tool and the approach of computational modelling with neural networks share the similarity that both are looking at brain functions at the level of *neuronal assembly*. In the case of EEG, at a scale of tens of thousands of pyramidal neurons; in the case of artificial neural networks, at a scale of networks of artificial neurons typically involving one or two hundreds such neurons.

I have always been fascinated by connectionist modelling. As I recalled the days when I first wrote programming code to build networks, I could hardly have anticipated that the networks I built could learn anything at all since the equations governing the dynamics of every single neuron merely allow them to perform simple associations.

Though modelling has often been lying on a fuzzy boundary between being *a toy* and being *an empirical investigation*, various experimentations with backpropagation networks as well as with simple recurrent networks have been keeping me curious about the original motivation of connectionist models of all kind, namely, to reflect how the actual brain might function.

In the later part of my study, an EEG facility has become accessible and it opens up the opportunity for me to study the real brain functioning for language.

# Part I

# Language Processing in Artificial Neural Networks

# Chapter 2

# Simple Recurrent Network Simulation 1 (SRN-SIM1)

## 2.1 Overview—Reassessing combinatorial productivity exhibited by simple recurrent networks

In this chapter a computational simulation on *reassessing* the ability of simple recurrent networks (SRNs) to generalise combinatorially will be reported.[1] van der Velde and colleagues (2004; 2005) proposed a framework and provided their assessment on SRNs' ability to generalise combinatorially. They attempted to demonstrate that:

(i) SRNs cannot generalise from training set sentences, such as "`the dog chased the cat`" and "`the boy saw the girl`", to sentences that contained new

[1] A portion of the findings discussed in this chapter has been reported in the *2006 International Joint Conference on Neural Networks (IJCNN)* (Wong, Minett & Wang, 2006)

combinations of lexical items, such as "the dog saw the cat" and "the boy chased the girl".

(ii) the reason for SRNs' failure to generalise is due to the networks' reliance on "word-word" association, i.e. the networks rely primarily on bi-gram statistics to process novel sentences.

The computational simulation reported in this chapter is a replication of the study of van der Velde et al. (2004) with the exception that a different network architecture was used.[2] The results I obtained were contrary to those of van der Velde et al. (2004). A better performance was attained by the networks and bi-gram statistics did not seem to be the dominant information the networks used to perform the task of prediction. In conclusion, I argue that the dismissal of SRNs as advocated by van der Velde was premature.

## 2.2 Combinatorial productivity and language acquisition

Compositionality is often regarded as the hallmark of human cognition (Fodor & Pylyshyn, 1988). It refers to events or objects where complex entities are composed by combining simple elements in a linear or hierarchical fashion. Language exhibits such a combinatorial nature at various levels, segments are combined to form syllables, syllables are combined to form words, words are combined to form sentences, etc.

[2] The implication of this will be discussed in Chapter 4

In the context of this chapter we consider a sentence to be a combination of lexical items. The complexity, as a consequence of such a combinatorial nature, in terms of number arises when we consider how one could learn to understand all possible combinations by just learning from a fraction of them, a scenario that all language learners face.

### 2.2.1 Combinatorial complexity in terms of number

To illustrate this, we *modelled* simple declarative sentences of English, e.g. "`the dog chased the cat`", as sequences of syntactic items taking, respectively, the subject, the verb and the object position. We denote such a sentence as an N-V-N sentence, "`n1-v2-n3`" being one such instance, to indicate that it involves a combination of members from the lexical categories of nouns (N) and verbs (V) to express a predicate-agent-patient meaning, i.e. the meaning of who ("`n1`") has done what ("`v2`") to whom ("`n3`").

With the assumption that N-V-N sentences are the only type of construction allowable in the language, $\mathcal{L}$. The size of the language, $|\mathcal{L}|$, i.e. the number of possible sentences in the language, grows with two quantities; one is the number of syntactic positions, $r$, to be filled in the construction, the other is the size of the two lexical categories, assuming both to be equal to $s$ for simplicity here. More importantly, the size of the language grows exponentially with $r$ and polynomially with $s$, as $|\mathcal{L}| = s^r$.

If we take a conservative estimate[3] that $\mathcal{L}$ permits only 1000 nouns and 1000 verbs. The total number of possible sentences in the language is $10^9$. If the language is to be

[3]Bloom (2000) gave an estimate of 60,000.

learned sentence-by-sentence, one needs more than 30 years to acquire the language even if he learns at a rate of one sentence per second.[4] Humans, however, exhibit *combinatorial productivity* (van der Velde et al., 2004; van der Velde & de Kamps, 2006), that is, the ability to generalise the knowledge of the language from limited samples to all possible sentences that contain never seen before combinations of lexical items. Such an ability to generalise enables the learner that once he has acquired the meaning of "*who has done what to whom*" from some N-V-N constructions, e.g. "`the dog chased the cat`" and "`the boy saw the girl`", he could automatically be able to comprehend all possible N-V-N constructions composed by the known lexicon, such as, among many others, "`the dog saw the cat`" and "`the boy chased the girl`".

### 2.2.2 Combinatorial productivity and connectionist networks

Van der Velde (2004; 2006) argued that connectionist networks, SRNs being one such "popular" type (Marcus, Vijayan, Rao & Vishton, 1999), lack the ability to generalise combinatorially and hence fail to capture an important feature of linguistic competence. The key element of their assessment on SRNs was the design of the training and testing set sentences in the simulations they reported.

I illustrate the rationale of the design with the two utterance networks shown in Figure 2.1 (p. 10). A sentence in $\mathcal{L}$ is represented by a path through the network from left to right. I consider $\mathcal{L}_C$ as a model of the language available to a child during his acquisition of the target adult language $\mathcal{L}_A$. $\mathcal{L}_C$ under-represents the target language in

[4] 60 sentences per minute × 60 (per hour) × 24 (per day) × 365 (per year) × 30 = $9.5 \times 10^8$

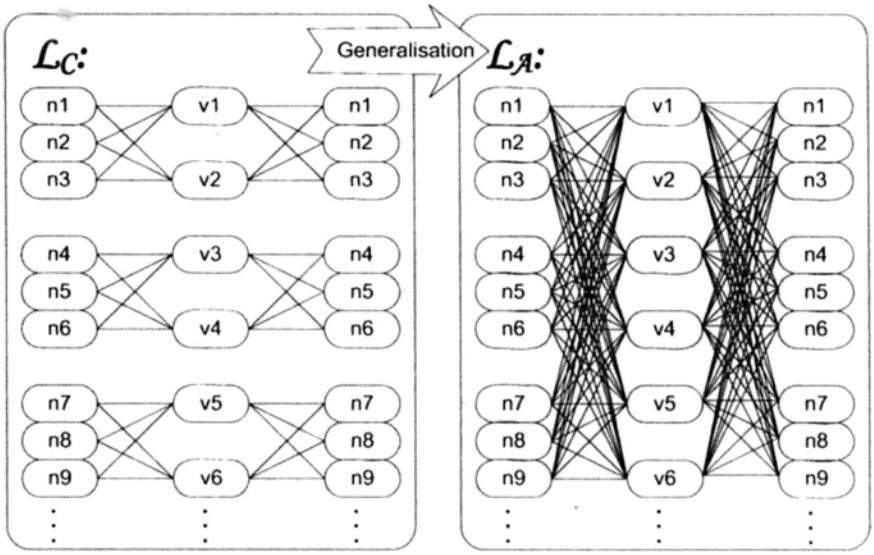



Figure 2.1: Combinatorial productivity as generalisation from the training set to the testing set. The training set, $\mathcal{L}_C$, is a model of the child directed speech that under-represents the target adult language, $\mathcal{L}_A$. In the simulation, SRNs were trained with sentences in $\mathcal{L}_C$ and their potential to exhibit combinatorial productivity was evaluated by their performance on processing novel sentences in $\mathcal{L}_A$.

a sense that most of the sentences in $\mathcal{L}_A$ are combinations of lexical items that are not in $\mathcal{L}_C$, e.g. "n1-v3-n7".

For SRNs to be a successful model of language acquisition, van der Velde et al. (2004) argued that SRNs should exhibit the ability to generalise combinatorially from $\mathcal{L}_C$ to $\mathcal{L}_A$. This ability can be evaluated by feeding only sentences from $\mathcal{L}_C$ to train the connection weights of the networks and testing networks' performance on both $\mathcal{L}_C$ and $\mathcal{L}_A$ sentences. If generalisation does occur, the networks' performance on $\mathcal{L}_A$ sentences should be comparable to that on $\mathcal{L}_C$ sentences. The training set and testing sets were thus constructed accordingly in van der Velde et al. (2004) and in the simulation reported here.

## 2.3 Methods

The connectionist architecture to be discussed in this study is the simple recurrent network (SRN) model which was proposed by Elman (1990) as a model for processing sequential information. It has evolved as models for the acquisition of syntax (Borovsky & Elman, 2006; Christiansen, Conway & Curtin, 2005; Elman, 2001) and sentence processing (Rodriguez, 2001; Christiansen & Chater, 1999b; Christiansen & Devlin, 1997).

The general architecture as well as the working mechanism of an SRN will be reviewed in Section 2.3.1 (p. 11). The section also describes the operations of the network implemented in the simulation to be reported in this chapter. Section 2.3.2 (p. 15) will describe the methods of evaluating network's performance regarding training and generalisation especially on the assessment of combinatorial productivity exhibited by SRNs. Section 2.3.3 (p. 21) will introduce the use of Grammatical Prediction Error (GPE) to evaluate SRNs' performance in leaning the language.

### 2.3.1 The Network architecture and the coding of the lexicon

Figure 2.2 (p. 12) shows the general architecture of SRNs that are commonly employed in the literature. Without the context layer, an SRN is just a layered feed-forward network in which every neuron in a layer is connected to every other neurons in the layer immediately above it. The component inside the box with dashed outline in Figure

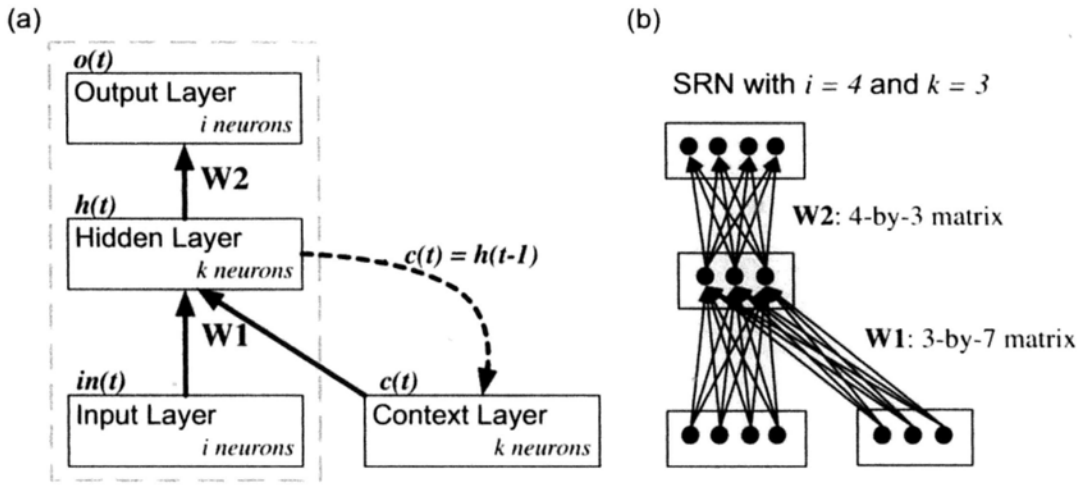



Figure 2.2: (a) The general architecture of a simple recurrent network employed in connectionist modelling of language processing. Arrows with solid lines denote full connection between layers of neurons, represented as blocks. Arrow with dotted line denotes the copy-back one-to-one connections. The directions of signal propagation during the feed-forward operation of an SRN are denoted by the arrows. (b) An example of SRN with 4 input and output neurons and 3 hidden layer neurons. **W1** and **W2**, the two weight matrices, denote the trainable connection weights between the layers of neurons.

2.2 (p. 12) denotes the feed-forward part of an SRN. The context layer in SRNs provides the network the ability to process sequential information through the accumulation of network's hidden layer activation. I will illustrate the working principle of an SRN with a scenario in which the network is to learn to process a sequence of words.

Suppose the network is to process the sequence ($v_1$, $v_2$, ...). The lexicon is by convention encoded with a set of orthogonal bit strings, where the $x^{\text{th}}$ bit of a bit string is set to 1 and the other bits are set to zero to encode the $x^{\text{th}}$ word. For example the bit string $\boldsymbol{v_1} = [1, 0, ..., 0]^{\text{T}}$ codes for the word $v_1$; $\boldsymbol{v_2} = [0, 1, ..., 0]^{\text{T}}$ codes for the word $v_2$, etc. The length of the bit code is equal to the number of input layer neurons, i.e. $i$.

At the first time step ($t = 1$), the input layer activation of the network, $\boldsymbol{in(t)}$, is

set to $\boldsymbol{v_1}$, the code for the first word in the sequence. Since $v_1$ is the first word of the sequence, the context layer activation, $\boldsymbol{c(t)}$, is set to a *null context* with an neutral value $[0.5, 0.5, ..., 0.5]^T$. The dimensions of $\boldsymbol{in(t)}$ and $\boldsymbol{c(t)}$ are, respectively, $i$ and $k$.

The concatenation of the vector $\boldsymbol{in(t)}$ and $\boldsymbol{c(t)}$ will then be fed to the network and propagated to the hidden layer via the weighted connection **W1**, a $k$-by-$(i + k)$ matrix, where $w1_{x,y}$ denotes the connection weight from the $y^{\text{th}}$ neuron in the input-context layer to the $x^{\text{th}}$ neuron in the hidden layer. The hidden layer activation , $\boldsymbol{h(t)}$, is given by the equation:

$$\boldsymbol{h(t)} = \varphi(\mathbf{W1}(\boldsymbol{in(t)} \oplus \boldsymbol{c(t)})) \tag{2.1}$$

where $\varphi$ denotes an activation function. In this study the logistic sigmoid function will be used where:

$$\varphi(x) = (1 + e^x)^{-1} \text{ and } \varphi([x_1, x_2, ...]) = [\varphi(x_1), \varphi(x_2), ...] \tag{2.2}$$

Similarly, the output of the network, i.e. the output layer activation, $\boldsymbol{o(t)}$, is obtained by propagating $\boldsymbol{h(t)}$ to the output layer via the connection weights between the hidden layer and the output layer, **W2**, i.e.

$$\boldsymbol{o(t)} = \varphi(\mathbf{W2}\ \boldsymbol{h(t)}) \tag{2.3}$$

The connection weights, **W1** and **W2**, are modified using the *backpropagation algorithm* (Haykin, 1999; Rumelhart, Hinton & Williams, 1986) with an objective to minimise the difference between actual output $\boldsymbol{o(t)}$ and a target output, $\boldsymbol{p(t)}$. Very often, a *prediction task* (Elman, 1990) is used to train the SRNs through which the networks are trained to associate the current word in a given context with the next word in a sequence.[5] Hence, $\boldsymbol{p(t)}$ is set to $\boldsymbol{v_{(t+1)}}$, the code for the next word in the sequence. This explains why $\boldsymbol{in(t)}$ and $\boldsymbol{o(t)}$ are of the same dimension. More about the prediction task and its use in this simulation will be discussed in page 15. Here I focus on giving a brief summary of the working mechanism of SRNs and provide notations for later discussion.

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 $\boldsymbol{c(t)}$ is set to $\boldsymbol{h(t-1)}$ which is the hidden layer activation of the network at the *previous time step*. The code for the second word in the sequence, $\boldsymbol{v_2}$ is now fed to the network as $\boldsymbol{in(t)}$ which is to be concatenated with $\boldsymbol{c(t)}$ and propagate to the hidden layer and output layer via Equations 2.1 and 2.3, respectively.

As the process of forward-feeding and backpropagation training continues through the iteration of the sequence, the context layer will keep track of the accumulating internal activation of the network. It is commonly denoted as one-to-one copy-back

[5]Elman (2004) provided three justifications for the use of prediction task, which I rephrase to suite the context of this thesis: (i) prediction forces the network to discover the underlying sentence structures that are manifested on the surface as linear sequences of strings; (ii) prediction training does not require explicit negative evidence as the success or failure of prediction can be verified on the fly; (iii) the use of prediction task does not assume that language learning is solely, nor mainly, about prediction, however, empirical evidence suggests that expectancy generation plays a role in language comprehension.

connection between the hidden layer and the context layer, the arrow with dotted line in Figure 2.2 (p. 12).

### 2.3.2 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 (Elman, 2004; Tomasello, 2003; Elman, 2001; Tomasello, 2001; Elman, 1999; Redington & Chater, 1998a; Redington, Chater & Finch, 1998) perspectives on plausible language acquisition mechanisms.

First, the usage-based nature of language learning (Tomasello, 2003; Tomasello, 2001) 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. This is exemplified by the attempt of connectionist models to provide an existence proof that compositional syntax can emerge out of simple associations between elements in sequences. To achieve this, SRNs are trained to associate the next word in the sequence with the current context. Historically, it is referred to as the *prediction task* since the output activation of the network after training reflects what word or set of words the network predicts will follow given the current word in a particular context.

I will illustrate the training of an SRN with a simple training set with only three sentences as shown in Table 2.1 (p. 16). The common practice in the SRN literature is

Table 2.1: A simple three-sentence-language to illustrate the training of a simple recurrent network

| To be fed to an SRN | The English equivalent |
|---|---|
| S1: "dog chased cat" | a simple declarative sentence—*the dog chased the cat* |
| S2: "dog barked" | a simple declarative sentence, with an intransitive main verb—*the dog barked* |
| S3: "cat that dog chased fled" | an object-extracted relative clause—*the cat that the dog chased fled* |

to concatenate the training set sentences into a single sequence of words. Each training set sentences usually appears in the sequence a number of times as if a child has heard the utterances a number of times. Also, the order of the sentences in the sequence is usually randomised in some way and each sentences are separated by an end-of-sentence marker, "#", modelling natural pauses between utterances. In the case of training an SRN with the training set sentences in Table 2.1 (p. 16), the sequence might take the form as in Table 2.2 (p. 16).

Table 2.2: A sequence generated from the three-sentenc-language (Table 2.1) to train a simple recurrent network on a prediction task

| |
|---|
| "dog chased cat # cat that dog chased fled # cat that dog chased fled # dog barked # dog chased cat # dog barked #... #" |

The sentences are fed to the network in a word-by-word manner following the order in the sequence. At the first time step, the sentence-initial word "dog" and a null context are fed to the network as the input. The network is trained to associate such an input with the next word in the sequence, i.e. "chased".[6] The outcome of this associative

[6]This is done via a coding scheme and mathematical operations on the coding, described in

learning is that, after training, when the network sees the word "dog" as the sentence-initial word, it will give an output activation that will better approximate the bit string that codes for the word "chased". The output of the network is often interpreted as the *prediction* the network has made about what words will follow the given context.

As subsequent words are fed to the network, its context layer keeps track of the sentence context through the accumulation of the information developed in the hidden layer of the network. SRNs can achieve that because the hidden layer activation at the previous time step is fed back to the network as the context layer activation, i.e. $\boldsymbol{c}(t) = \boldsymbol{h}(t-1)$. This context layer activation is part of the input to the network and contributes to the value of the hidden layer activation at the current time step via the connection weights between the context layer and the hidden layer of the network, and it will in turn contribute to the hidden layer activation at the next time step. Table 2.3 (p. 18) lists the information flow at each time step during the training of an SRN with the sequence in Table 2.2 (p. 16).

A 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. This assessment is based on the coding scheme and the prediction training described above. Recall that the lexicon is encoded with a set of orthogonal bit strings, where the $x^{\text{th}}$ bit of a bit string is set to 1 and the other bits are set to 0 to encode word $x$. The coding scheme for the three-sentnce-language is given in Table 2.4 (p. 19). Because

Section 2.3.1 (p. 11)

Table 2.3: Information flow during the training of an SRN with the sequence in Table 2.2. The figure below is a simplified notation of an SRN highlighting via which layer of neurons a word will be fed to the network. At each time step, backpropagation algorithm is applied to minimise the difference between the actual output activation of the network $o(t)$ and the target output activation $p(t)$.

| Time step ($t$) | Context ($c(t)$) | Current word ($in(t)$) | Target output ($p(t)$) |
|---|---|---|---|
| 1 | null | dog | chased |
| 2 | dog | chased | cat |
| 3 | dog chased | cat | # |
| 4 | dog chased cat | # | cat |
| 5 | ... # | cat | that |
| 6 | ... # cat | that | dog |
| 7 | ... cat that | dog | chased |
| 8 | ... cat that dog | chased | fled |
| 9 | ... cat that dog chased | fled | # |
| 10 | ... cat that dog chased fled | # | cat |
| 11 | ... # | cat | that |
| ⋮ | ⋮ | ⋮ | ⋮ |

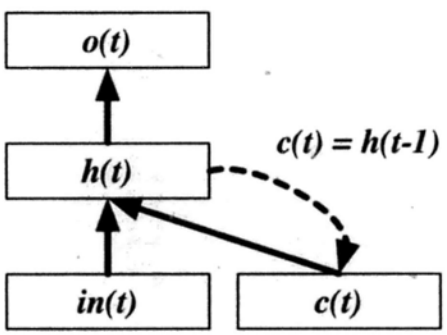


Table 2.4: The codebook for the lexicon of the three-sentence-language

| | Bit | | | | | | |
|---|---|---|---|---|---|---|---|
| **Word** | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th |
| dog | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| cat | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| chased | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| barked | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| fled | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| that | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| # | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

the SRN is trained with the prediction task where the output layer activation[7] is tuned to approximate the bit string coding for the word that immediately follows the current context, the output layer activation can thus be interpreted as the network's estimate of the probability distribution regarding which words would follow a given context.

Another way to interpret the network's output is to treat the output layer as if it is coding the lexicon using a one-word-one-neuron scheme. This is illustrated in Figure 2.3 (p. 20) where an SRN that has been trained with the three-sentence-language (Table 2.1, p. 16) is given a sentence-initial word "dog" as the input. If the network has acquired the language it would give an output layer activation in which the third and the fourth output layer neuron would be most active indicating that both the word "chased" and "barked" are equally probable in the context of a sentence-initial word "dog".

It is important to note that bi-gram statistics alone are insufficient for the network

[7]Output layer activation, $\boldsymbol{o}(t) = \varphi(\mathbf{W2}\ \boldsymbol{h}(t))$ (Equation 2.3, p. 13), is a vector with the same dimension as the codebook.

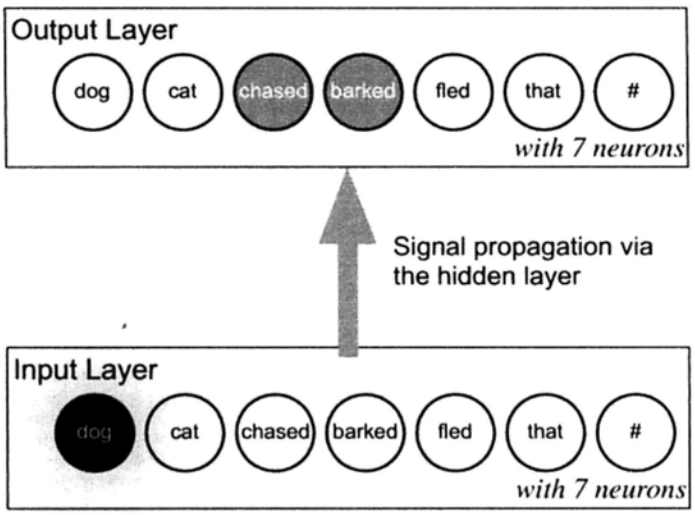



Figure 2.3: A sentence-initial word "dog" is fed to an SRN trained with the three-sentence-language (Table 2.1, p. 16). The darkness of the node indicates the activation value of a neuron, black: 1, white: 0, grey: 0.5. For simplicity, only the input and output layer are shown.

to make grammatically correct prediction. This is evidenced by the processing of the object-extracted relative clause[8] where bi-gram statistics would suggest that both the word "chased" and "barked" could follow the word "dog" in the context "cat that dog ...". The prediction based on bi-gram that "barked" could follow is ungrammatical according to the language since "barked" is an intransitive verb. To overcome such a false alarm, the network has to be sensitive to the sentence context in which a word apprears, Figure 2.4 (p. 21) illustrates a scenario where the context layer of an SRN is in action allowing the network to make the correct prediction that the output layer neuron for the word "barked" should not be activated.

[8] "cat that dog chased fled", S3 in the three-sentence-language.

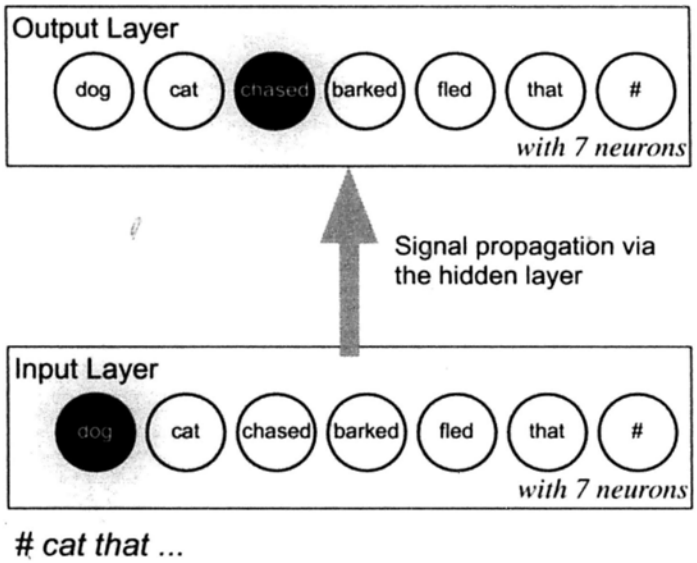



Figure 2.4: The word "dog" in context of the sentence "`cat that dog chase fled`" (S3 in Table 2.1, p. 16) is fed to the SRN. The darkness of the node indicates the activation value of a neuron, black: 1, white: 0. For simplicity, only the input and output layer are shown. $GPE = 0$.

### 2.3.3 Grammatical Prediction Error (GPE)

I have illustrated in the last section the prediction training of an SRN and the way to interpret the network's output layer activation in terms of the grammatical correctness of the network's prediction about what words could follow in a given sentence context. I now continue with the discussion on the assessment of SRNs' ability to exhibit combinatorial productivity of language.

In the simulation of van der Velde et al. (2004), they evaluated the success of the networks in processing the language with Grammatical Prediction Error (GPE).[9] In short, GPE is a measurement of the grammaticality of an SRN's prediction. For each sentence position of each sentence construction, we can define, according to

[9]But see (Christiansen & Chater, 1999a) for alternative definition of GPE.

the language, a set of words that are considered to be the correct continuations and at the same time a set words that are the incorrect continuations. To achieve high grammaticality, the network should activate those correct words and suppress the activity to words that are incorrect. The definition of GPE reflects such a requirement:

$$GPE = \frac{\sum \text{Incorrect Activation}}{\sum \text{Correct Activation} + \sum \text{Incorrect Activation}} \tag{2.4}$$

where the numerator is the sum of the activations of the output neurons coding for words that are grammatically incorrect continuations and the first part of the denominator is the sum of the activations of the output neurons coding for words that are grammatically correct continuations.

In the case of an SRN that has acquired the three-sentence-language and gives an output layer activation as shown in Figure 2.4 (p. 21) when the word "dog" is fed as the input in the context of the S3 sentence "cat that dog chased fled", the network achieved a GPE of zero since nodes other than the only correct continuation "chased" are not activated. On the other hand, if an SRN fails to capture the sentence context and relies on bi-gram statistics to make the prediction, it will wrongly activate the node for the intransitive verb "barked" results in a hight GPE evaluation of 0.5, as illustrated in Figure 2.5 (p. 23)

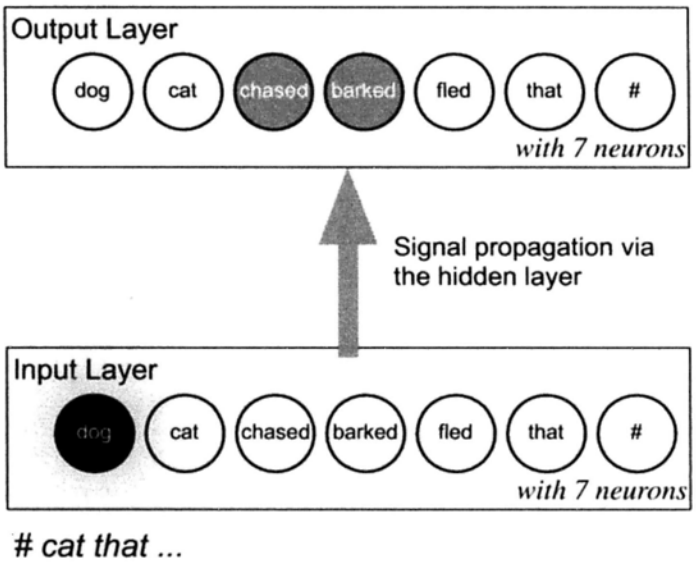



Figure 2.5: The word "dog" in context of the sentence "cat that dog chase fled" (S3 in Table 2.1, p. 16) is fed to the SRN. The darkness of the node indicates the activation value of a neuron, black: 1, white: 0, grey: 0.5. For simplicity, only the input and output layer are shown. $GPE = 0.5$.

## 2.4 Materials

Having introduced the use of the simple recurrent networks to model sentence processing, I now turn to the focus of the simulation of this chapter, namely, to look at combinatorial productivity exhibited by SRNs. The framework of assessment was introduced by van der Velde et al. (2004) 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 and language processing. The key element of the study was the design of the training and the testing set sentences. The testing set contained all possible combinations of lexical items in composing several sentence constructions whereas the training set contained only a fraction of those combinations. The idea is introduced in Section 2.2.2 (p. 9) as

Table 2.5: Three types of sentence construction used in the simulation

| **Sentence type** | **Construction** | **English equivalent** |
|---|---|---|
| Simple | N-V-N-# | The dog chased the cat. |
| Right-branching | N-V-N-that-V-N-# | The dog chased the cat that caught the mouse. |
| Centre-embedding | N-that-N-V-V-N-# | The cat that the dog chased caught the mouse. |

graphically illustrated in Figure 2.1 (p. 10).

In this section I will report the actual implementation of the training set and testing sets in my replication of the simulation experiment of van der Velde et al. (2004).

### 2.4.1 The training and the testing sets

The networks were trained with three types of sentences, simple, right-branching and centre-embedding sentences, as tabulated in Table 2.5 (p. 24). 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 lexicon of nouns and verbs was divided into four non-overlapping groups, I denote the $j^{\text{th}}$ member of the $i^{\text{th}}$ group of nouns as "$\mathsf{n}_{ij}$" and similarly "$\mathsf{v}_{ij}$" for verbs, as illustrated in Table 2.6 (p. 25). Together with two dummy

Table 2.6: The lexicon of nouns and verbs used in the simulation

| Group | Noun | Verb |
|---|---|---|
| 1 | $n_{11}, n_{12}$ | $v_{11}, v_{12}$ |
| 2 | $n_{21}, n_{22}$ | $v_{21}, v_{22}$ |
| 3 | $n_{31}, n_{32}$ | $v_{31}, v_{32}$ |
| 4 | $n_{41}, n_{42}$ | $v_{41}, v_{42}$ |

words "$d_{01}$" and "$d_{02}$", there were 20 words in the lexicon and they were coded with an 1-in-20-bit coding scheme as show in Table 2.7 (p. 26). The dummy words were not used in the simulation, their presence was for a compatibility reason as van der Velde et al. (2004) made use of SRNs with 20 input and output layer neurons.

Under the framework of van der Velde et al. (2004) and the way the networks' performance is evaluated with GPE, I consider that only the nouns and the verbs were essential word classes to be modelled because the major hypothesis to be tested in the simulation was whether the ability to generalise by SRNs is solely determined by bi-gram statistics. The inclusion of the relative marker "`that`" was mainly for the purpose of compatibility with the original study of van der Velde et al. (2004).

**The training set**

Training set sentences were composed of nouns and verbs from the ***same*** group. Eight unique sentences ($2^3$, 2 different words at 3 different syntactic positions) were generated for each group of the simple sentence. Similarly 32 unique sentences ($2^5$, 2 different words at 5 different syntactic positions) were generated for each group of the right-

Table 2.7: The coding scheme of the lexicon in the simulation

| **Word** | **1st to 20th Bit** | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_{11}$ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n_{12}$ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n_{21}$ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n_{22}$ | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n_{31}$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n_{32}$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n_{41}$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n_{42}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $v_{11}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $v_{12}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $v_{21}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $v_{22}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $v_{31}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $v_{32}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| $v_{41}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| $v_{42}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| that | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| # | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $d_{01}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $d_{02}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Table 2.8: The construction of training set sentences

| Sentence type | Group | The set | Set size |
|---|---|---|---|
| Simple | 1 | $\{n_{1a}\text{-}v_{1b}\text{-}n_{1c}\text{-}\#\}$ | 8 |
| | 2 | $\{n_{2a}\text{-}v_{2b}\text{-}n_{2c}\text{-}\#\}$ | 8 |
| | 3 | $\{n_{3a}\text{-}v_{3b}\text{-}n_{3c}\text{-}\#\}$ | 8 |
| | 4 | $\{n_{4a}\text{-}v_{4b}\text{-}n_{4c}\text{-}\#\}$ | 8 |
| Right-branching | 1 | $\{n_{1a}\text{-}v_{1b}\text{-}n_{1c}\text{-that-}v_{1d}\text{-}n_{1e}\text{-}\#\}$ | 32 |
| | 2 | $\{n_{2a}\text{-}v_{2b}\text{-}n_{2c}\text{-that-}v_{2d}\text{-}n_{2e}\text{-}\#\}$ | 32 |
| | 3 | $\{n_{3a}\text{-}v_{3b}\text{-}n_{3c}\text{-that-}v_{3d}\text{-}n_{3e}\text{-}\#\}$ | 32 |
| | 4 | $\{n_{4a}\text{-}v_{4b}\text{-}n_{4c}\text{-that-}v_{4d}\text{-}n_{4e}\text{-}\#\}$ | 32 |
| Centre-embedding | 1 | $\{n_{1a}\text{-that-}n_{1b}\text{-}v_{1c}\text{-}v_{1d}\text{-}n_{1e}\text{-}\#\}$ | 32 |
| | 2 | $\{n_{2a}\text{-that-}n_{2b}\text{-}v_{2c}\text{-}v_{2d}\text{-}n_{2e}\text{-}\#\}$ | 32 |
| | 3 | $\{n_{3a}\text{-that-}n_{3b}\text{-}v_{3c}\text{-}v_{3d}\text{-}n_{3e}\text{-}\#\}$ | 32 |
| | 4 | $\{n_{4a}\text{-that-}n_{4b}\text{-}v_{4c}\text{-}v_{4d}\text{-}n_{4e}\text{-}\#\}$ | 32 |

where $a, b, c, d = \{1, 2\}$

branching as well as the centre-embedding sentences. The full set of 288 training set sentences are listed in Table 2.8 (p. 27).

The four groups of sentences were combined to form the training sets with different ratios of simple, right-branching and centre-embedding sentences according to the four-phase training scheme. In Phase I, 32,000 sentences were sampled randomly without replacement, i.e. each sentence was sampled 1,000 times, from the set of group 1 to group 4 sentences and fed to the SRNs for the prediction training. Starting from Phase II, complex sentences (right-branching and centre-embedding) started to be included in the training data. Table 2.9 (p. 28) lists the total number of sentences fed to a network during each phase of the training process.

The design of the training scheme with increasing number of complex sentences was in accordance with Elman's notion of "starting small" (Elman, 1993; van der

Table 2.9: The four-phase training scheme for SRN-SIM1

| | | Number sentences fed to a network | | | |
|---|---|---|---|---|---|
| **Phase** | **Token ratio**† | *Simple* | *Right-branching* | *Centre-embedding* | *Total* |
| I | 1 : 0 : 0 | 32,000 | 0 | 0 | 32,000 |
| II | 6 : 1 : 1 | 7,680 | 1,280 | 1,280 | 10,240 |
| III | 2 : 1 : 1 | 25,600 | 12,800 | 12,800 | 51,200 |
| IV | 1 : 2 : 2 | 12,800 | 25,600 | 25,600 | 64,000 |
| **Total** | | 78,080 | 39,680 | 39,680 | 157,440 |

†Ratio of simple : right-branching : centre-embedding sentences

Velde et al., 2004), my 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 with testing set sentences, via GPE as introduced in Section 2.3.3 (p. 21).

**The testing sets**

Recall that the objective of the simulation is to evaluate the combinatorial productivity exhibited by SRNs. The test set sentences were constructed such that they involved combination of lexical items that the networks had not been exposed to before. To achieve that, testing set sentences were constructed by combining lexical items from ***mixed groups***. The level of difficulty with respect to generalisation was varied by the number of groups that were mixed. The more the number of groups the more difficult the sentence would be. I use $M$, the number of groups of lexical items from which a sentence was constructed, to denote such a level of complexity of a testing set sentence.

As the lexical items of nouns and verbs were divided into four groups, for each sentence type, there were a total of ${}_4P_2 = 12$ subsets of $M = 2$ testing set sentences.[10] And the size of each subset varied depending on the sentence length, 8 for simple sentences and 32 for complex sentences. Similarly, there were ${}_4P_3 = 24$ subsets of $M = 3$ and ${}_4P_4 = 24$ subsets of $M = 4$ testing set sentences, respectively. The construction of the testing sets sentences is summarised in Table 2.10 (p. 30). Obviously, training set sentences were with M-value of 1.

### 2.4.2 Network architecture

Twenty SRNs with the architecture as shown in Figure 2.6 (p. 31) were employed in the simulation. Each network was initialised with an independent random set of connection weights with values drawn from a normal distribution with a mean of zero and a standard deviation of 0.05. The learning rate and momentum was set to 0.1 and 0, respectively. The networks were trained with the prediction task with streams of concatenated sentences which were randomly sampled from the training sets. The results to be reported in the remaining parts of the chapter were based on the average performance of the twenty networks.

## 2.5 Results

After the fourth phase of training, the connection weights of the networks were frozen. Testing set sentences of various types and M-values[11] were fed to the networks through

[10] ${}_nP_r$ is the number of permutations of selecting $r$ elements from a set of size $n$. ${}_nP_r = \frac{n!}{(n-r)!}$

[11] $M$ is the number of groups of lexical items that were mixed to construct the testing sets

Table 2.10: The construction of testing sets sentences

| Sentence type | M-value | Example subset | Groups mixed | No. of subsets | Total set size |
|---|---|---|---|---|---|
| Simple | 2 | $\{n_{1a}\text{-}v_{3b}\text{-}n_{1c}\text{-}\#\}$<br>$\{n_{2a}\text{-}v_{4b}\text{-}n_{2c}\text{-}\#\}$ | 1 and 3<br>2 and 4 | ${}_4P_2 = 12$ | 96 |
| | 3 | $\{n_{1a}\text{-}v_{2b}\text{-}n_{3c}\text{-}\#\}$<br>$\{n_{4a}\text{-}v_{2b}\text{-}n_{1c}\text{-}\#\}$ | 1, 2 and 3<br>4, 2 and 1 | ${}_4P_3 = 24$ | 192 |
| Right-branching | 2 | $\{n_{1a}\text{-}v_{3b}\text{-}n_{1c}\text{-that-}v_{3d}\text{-}n_{1e}\text{-}\#\}$<br>$\{n_{2a}\text{-}v_{4b}\text{-}n_{2c}\text{-that-}v_{4d}\text{-}n_{2e}\text{-}\#\}$ | 1 and 3<br>2 and 4 | ${}_4P_2 = 12$ | 384 |
| | 3 | $\{n_{1a}\text{-}v_{2b}\text{-}n_{3c}\text{-that-}v_{1d}\text{-}n_{2e}\text{-}\#\}$<br>$\{n_{4a}\text{-}v_{2b}\text{-}n_{1c}\text{-that-}v_{4d}\text{-}n_{2e}\text{-}\#\}$ | 1, 2 and 3<br>4, 2 and 1 | ${}_4P_3 = 24$ | 768 |
| | 4 | $\{n_{1a}\text{-}v_{2b}\text{-}n_{3c}\text{-that-}v_{4d}\text{-}n_{1e}\text{-}\#\}$<br>$\{n_{4a}\text{-}v_{2b}\text{-}n_{1c}\text{-that-}v_{3d}\text{-}n_{4e}\text{-}\#\}$ | 1, 2, 3 and 4<br>4, 2, 1 and 3 | ${}_4P_4 = 24$ | 768 |
| Centre-embedding | 2 | $\{n_{1a}\text{-that-}n_{3b}\text{-}v_{1c}\text{-}v_{3d}\text{-}n_{1e}\text{-}\#\}$<br>$\{n_{2a}\text{-that-}n_{4b}\text{-}v_{2c}\text{-}v_{4d}\text{-}n_{2e}\text{-}\#\}$ | 1 and 3<br>2 and 4 | ${}_4P_2 = 12$ | 384 |
| | 3 | $\{n_{1a}\text{-that-}n_{2b}\text{-}v_{3c}\text{-}v_{1d}\text{-}n_{2e}\text{-}\#\}$<br>$\{n_{4a}\text{-that-}n_{2b}\text{-}v_{1c}\text{-}v_{4d}\text{-}n_{2e}\text{-}\#\}$ | 1, 2 and 3<br>4, 2 and 1 | ${}_4P_3 = 24$ | 768 |
| | 4 | $\{n_{1a}\text{-that-}n_{2b}\text{-}v_{3c}\text{-}v_{4d}\text{-}n_{1e}\text{-}\#\}$<br>$\{n_{4a}\text{-that-}n_{2b}\text{-}v_{1c}\text{-}v_{3d}\text{-}n_{4e}\text{-}\#\}$ | 1, 2, 3 and 4<br>4, 2, 1 and 3 | ${}_4P_4 = 24$ | 768 |

where $a, b, c, d = \{1, 2\}$

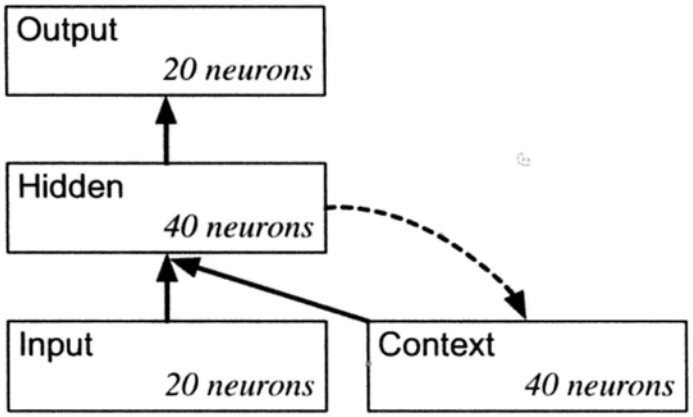



Figure 2.6: Simple recurrent network architecture used in the simulation

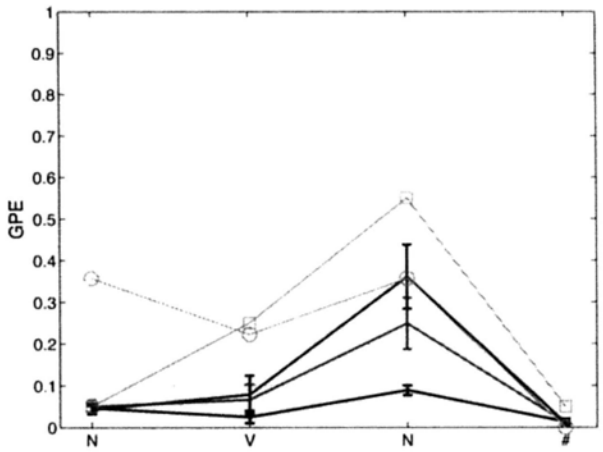
which the output activations of the networks were subjected to GPE evaluation on grammaticality.

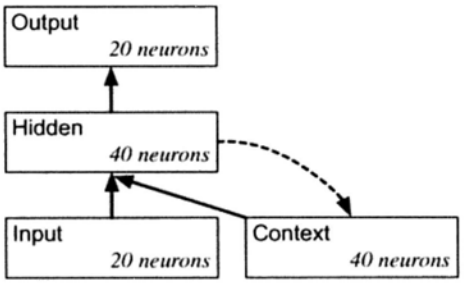
Each network was evaluated with 100 sentences drawn randomly from each subset, sentence type by M-value,[12] of the testing set sentences. The mean GPEs attained by each SRN at each sentence position were recorded and the grand mean GPEs, averaged across all networks, were plotted in Figure 2.7 (p. 32). Except for the lines in grey colour, a data point in the figure represents the GPEs of SRNs' outputs in predicting the next lexical item at the sentence position indicated on the $x$-axis. For example, the first data point of each graph in Figure 2.7 (p. 32) represents the grand mean GPE of networks' outputs given the sentence-inital nouns were fed as the inputs and at this sentence position either a verb or the relative marker "`that`" was considered to be the grammatically correct prediction.

The GPE evaluations on various subsets of testing set sentences, $M = 2$ (green), $M = 3$ (blue) and $M = 4$ (red), are to be compared with some reference values:

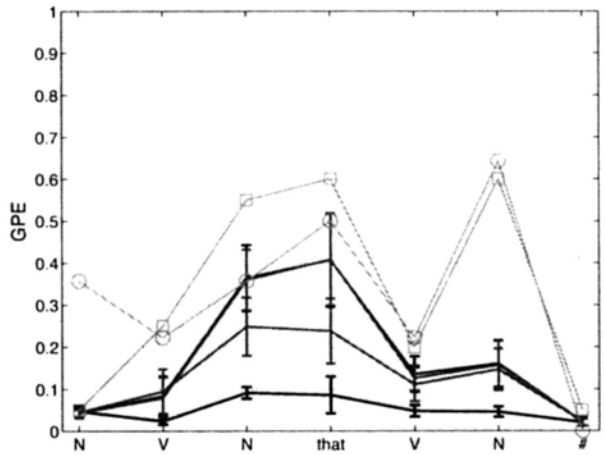
[12] cf. The construction of testing sets in Table 2.10 (p. 30).

Simple




Right-branching




Centre-embedding

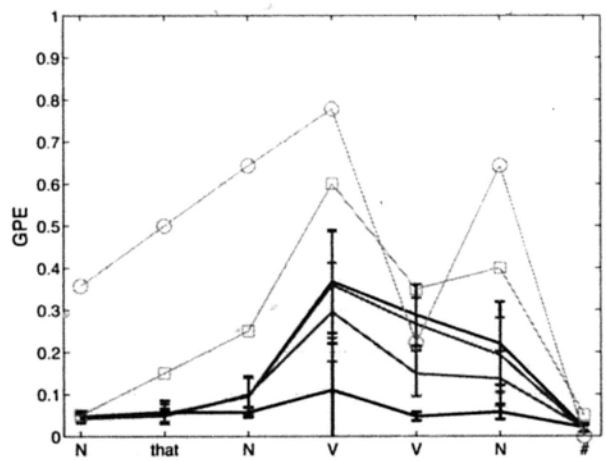




Figure 2.7: GPE evaluation on the training and testing set sentences of simple (top), right-branching (middle) and centre-embedding (bottom) sentences. Black: GPEs on training set sentences; green: GPEs on $M = 2$ testing set sentences; blue: $M = 3$; red: $M = 4$. Grey lines with square marks: GPEs on $M = 3$ testing set sentence in the simulation of van der Velde et al. (2004). Grey lines with circle marks: expected GPEs by a bi-gram model. The error bars indicate plus and minus one standard deviation from mean GPEs of the networks I trained.

Table 2.11: The bi-gram transitions. The value of the cell in the $i^{th}$ row $j^{th}$ column is the probability that the words in the category $j$ follow the words in category $i$. Formally, $Pr(w_{k+1} \in \mathbf{C}_j | w_k \in \mathbf{C}_i)$, where $w_k$ and $w_{k+1}$ are consecutive words in a sequence.

| | **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 |

- Training set GPEs (M=1), plotted in black line
- GPEs on testing sets reported in van der Velde et al. (2004), which I mark on Figure 2.7 (p. 32) with grey line with square line markers, labelled "VDV M=3"
- Expected GPEs obtained by a bi-gram model, grey line with circle line markers. The bi-gram model was constructed from the empirical transition probabilities (available from the training data) among the four lexical categories, N, V, that and # during the last phase of training. The bi-gram statistics were tabulated in Table 2.11 (p. 33).

A general trend that can be observed from Figure 2.7 (p. 32) is that the SRNs learned the training set sentences nearly perfectly, with very low GPEs, but performed worse and worse on testing sets as the complexity in terms of generalisation increased from $M = 2$ to $M = 4$. This was consistent with the original experiment reported by van der Velde et al. (2004). As the main focus of the simulation was to evaluate how well SRN generalise combinatorially, I will focus on the performance of the networks on testing set sentences.

The most apparent difference between my data and those of van der Velde et al. (2004) was that at all sentence positions the SRNs in my simulation attained a lower mean GPEs than those in van der Velde's. The improvement was also significant at the sentence positions that were expected to be difficult, such as at the sentence final nouns and at the first verbs of centre-embedding sentences. Notice that, for the two complex sentences, an $M = 3$ testing set was as complex as an $M = 4$ testing set until the second verb was encountered because the partial sentences up to this position involved only three content words. This explains why the two sets of GPE plots overlap with one another for the first four sentence positions.

This simulation did not intend to provide evidence that SRNs can exhibit full generalisation capability that testing sets GPEs can be as low as the training set GPEs. Rather, it aimed at providing evidence against the major claim of van der Velde et al. (2004) that when an SRN fails to generalize it resorts to word-word association between immediately adjacent words, i.e. the bi-gram statistics, in which they put:

> "*... when words from different lexicon groups are mixed, the networks do not predict the next word (lexical category) on the basis of the sentence context, but primarily on the basis of the associations between the words learned during training.*"
>
> van der Velde et al. (2004, p. 34)

Their argument was based on the observation that the testing set GPEs they obtained followed the trend of a bi-gram GPEs. The GPE evaluation of my simulation suggested that the role bi-gram statistics in SRN was overstated.

In my simulation, although the GPEs in general followed a similar pattern with van der Velde's, the GPEs of my networks attained smaller values in all sentence positions. In those positions that they argued as being the most difficult to make correct predictions, such as at the last word of all three sentence types and at the first verb of a centre-embedding sentence, not only the improvements were substantial, the GPEs attained in my simulation were much smaller than the GPEs a bi-gram model could achieve. In particular, at the first verb of a centre-embedding sentence, the frequent V-N transitions strongly suggest[13] a noun should be predicted, yet my networks attained a much lower GPE of about 0.36 on $M = 4$ testing set sentences than the expected GPE of 0.78 if SRNs relied primarily on bi-gram statistics.

Among the fifteen sentence positions, excluding the end-of-sentence marker, at only four of them the networks showed a mean GPE approximately equal to or larger than a bi-gram GPE. These four positions were: the sentence final noun of simple sentences, the second noun and the relative marker of right-branching sentences and the second verb of centre-embedding sentences. To obtain a clearer picture about the behaviour of the networks, the analysis has to go beyond the surface GPE scores and has to take into account the activation patterns of the output layer neurons during the processing of a sentences.

[13] cf. Table 2.11 (p. 33), the bi-gram statistics, with a probability of 0.778 for a V-N transition and 0.222 for a V-V transition, and hence an expected bi-gram GPE of 0.778 at the first verb of a centre-embedding sentence.

## 2.6 Analysis of networks' output layer activation

One $M = 4$ testing set right-branching sentence[14] and one $M = 4$ centre-embedding sentence[15] were sampled randomly for the analysis of networks' output layer activation. Each network was fed with the two sentences and their output layer activations at each sentence position were recorded. The output layer activations averaged across the twenty SRNs were plotted in Figure 2.8 (p. 37). In the figure, each bar chart represents the average output layer activation when the word indicated on the left was fed to the networks. The $y$-axis denotes the activation of an output layer neuron coding for a word indicated on the $x$-axis.

For instance, in penal (a) of Figure 2.8, the first (from the top) bar chart represents the output layer activation when the sentence-initial noun "$n_{42}$" of the right-branching sentence was fed to the networks. The networks were predicting that the two group 4 verb or the relative marker "that" should follow and the prediction was indeed grammatical and hence a low GPE was achieved. The second bar chart represents the output layer activation when the word "$v_{11}$" was fed as the second word of the sentence, similarly the time course of output layer activations for the remaining words of the right-branching sentence were plotted in the other bar charts of penal (a). Likewise penal (b) of Figure 2.8 shows the time course of output layer activations of the networks during the processing of the centre-embedding sentence.

[14]The sampled right-branching sentence was: "$n_{42}$-$v_{11}$-$n_{22}$-that-$v_{31}$-$n_{41}$-#"
[15]The sampled centre-embedding sentence was: "$n_{31}$-that-$n_{22}$-$v_{41}$-$v_{12}$-$n_{31}$-#"

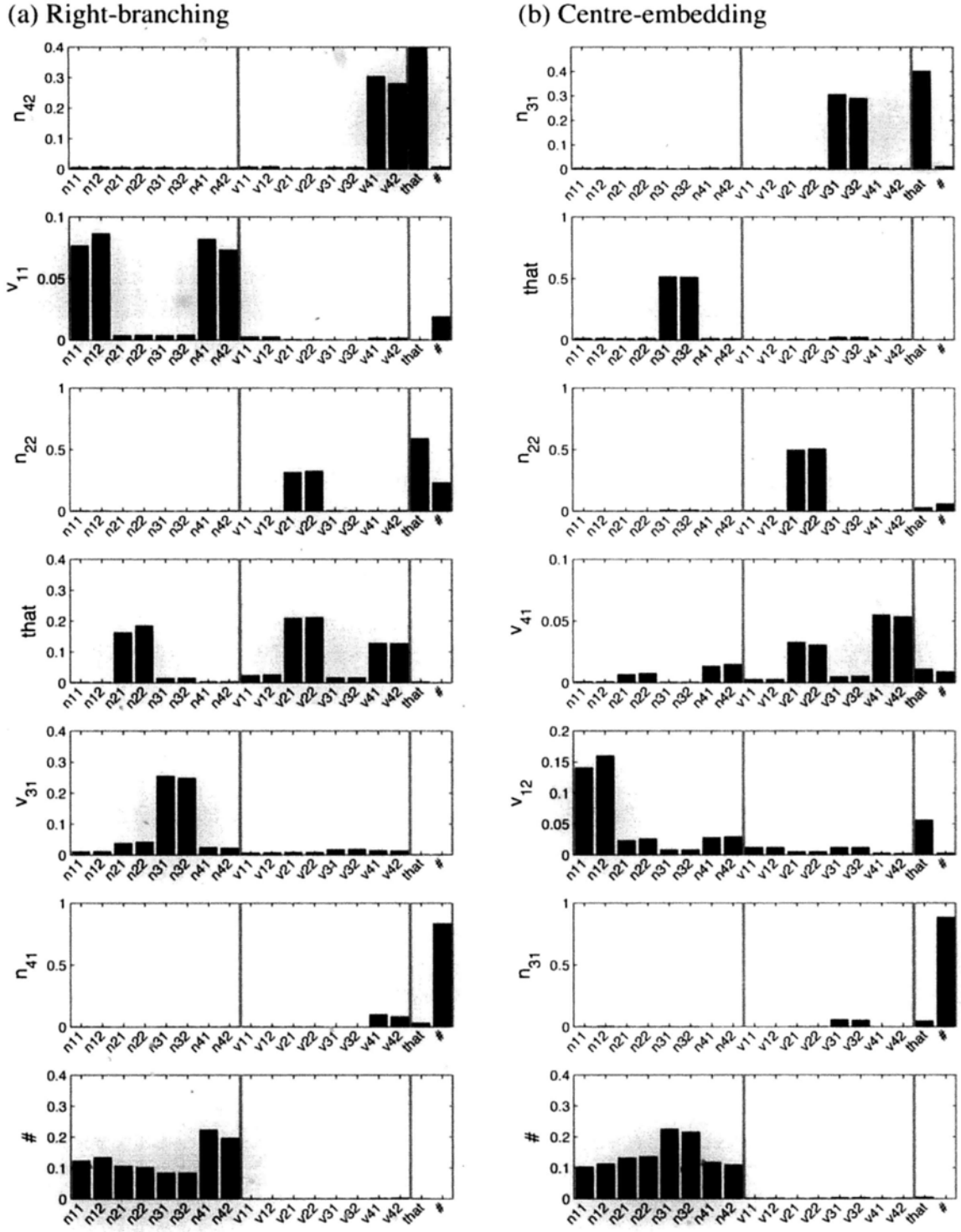



Figure 2.8: The time course of output layer activations of the twenty SRNs during the processing of (a) an $M = 4$ right-branching sentence and (b) an $M = 4$ centre-embedding sentence. See text in Section 2.6 (p. 36) for details.

I have mentioned that there were four sentence positions where the GPE attained by the networks were approximately equal to or larger than a bi-gram GPE. They were:

- The sentence final noun of simple sentences
- The second noun of right-branching sentences
- The relative marker "that" of right-branching sentences
- The second verb of centre-embedding sentences

Based on such an observation of a *bi-gram trend*, van der Velde et al. (2004) argued that SRN relies primarily on word-word association to perform the task of prediction. However, they did not examine the output activations of the networks to further justify their claim. In the following, by examining the output activations at some critical positions where SRN seems to be having the greatest problem in processing the language, I argue for a weaker version of the claim of van der Velde et al. (2004).

### 2.6.1 The sentence final noun of simple sentences and the second noun of right-branching sentences

At the second noun of the right-branching sentence, the third bar chart in penal (a) of Figure 2.8 (reproduced as Figure 2.9, p. 39), the networks gave an average output layer activation in which the node coding for the relative marker "that" was activated the strongest. The two group 2 verbs, "$v_{21}$" and "$v_{22}$", as well as the end-of-sentence

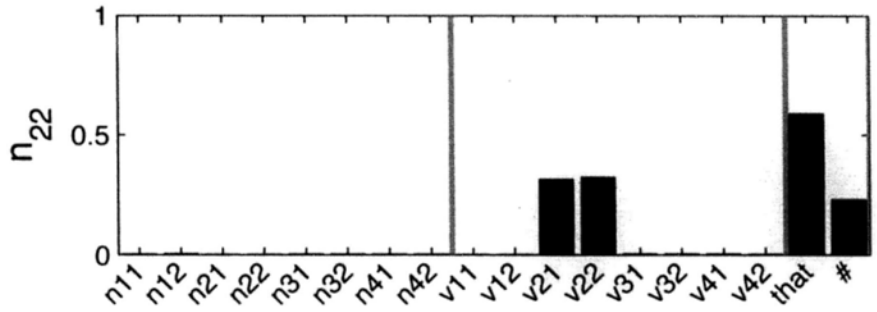



Figure 2.9: Output layer activation at "$n_{42}$-$v_{11}$-$n_{22}$-..." during the processing of a right-branching sentence. The third bar chart in penal (a) of Figure 2.8.

marker "#" also received high activation. This indicates that the networks were indeed having difficulties processing sentences with novel combinations of nouns and verbs from mixed groups because their predictions that verbs could follow the partial sentence "$n_{42}$-$v_{11}$-$n_{22}$-..." were grammatically incorrect.

### 2.6.2 The relative marker "that" of right-branching sentences

As for another position where a bi-gram-like GPE was obtained, at the "that" slot of right-branching sentences, the grammatically correct continuation comprised verbs only while bi-gram statistics (the third row of Table 2.11, p. 33) suggested that nouns and verbs were equally probable continuations.

The activation patterns at this sentence position were sampled and plotted in the fourth bar chart in penal (a) of Figure 2.8 (reproduced as Figure 2.10, p. 40). As can be observed from the figure, contrary to bi-gram statistics, the networks gave twice as much activation to verbs than to nouns output neurons. Notice that, not only did the networks gave the two group 2 verbs, "$v_{21}$" and "$v_{22}$", high activation, the two

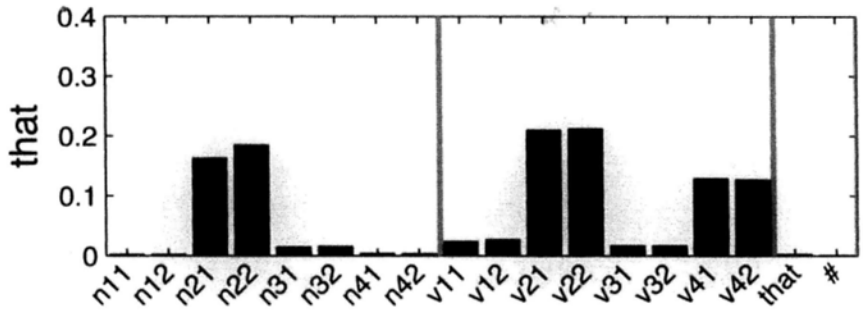



Figure 2.10: Output layer activation at "$n_{42}$-$v_{11}$-$n_{22}$-that-..." during the processing of a right-branching sentence. The fourth bar chart in penal (a) of Figure 2.8.

group 4 verbs , "$v_{41}$" and "$v_{42}$", were also activated at a considerable degree. More importantly, the activation of the group 4 verbs was evidence to indicate that SRNs retained the ability to keep track of the novel sentences because the activation reflected the networks' memory of a group 4 noun, "$n_{42}$", that had been fed as the sentence-initial word. However, to exhibit human-level combinatorial productivity towards sentence processing, one might expect that all verbs should be activated, in this simulation there was no sign of this level of productivity.

### 2.6.3 The second verb of centre-embedding sentences

The average output activation at the second verb of the centre-embedding sentence was plotted in the fifth bar chart in penal (b) of Figure 2.8 (reproduced as Figure 2.11, p. 41). At this sentence position, though grammatically incorrect, the networks were showing activation for the relative marker "that" which was beyond the expectation of a bi-gram model. The networks might have wrongly segmented the testing set sentence and treated the preceding two words as the first and second word therefore made the prediction that "that" should appear based on absolute sentence position. In the other

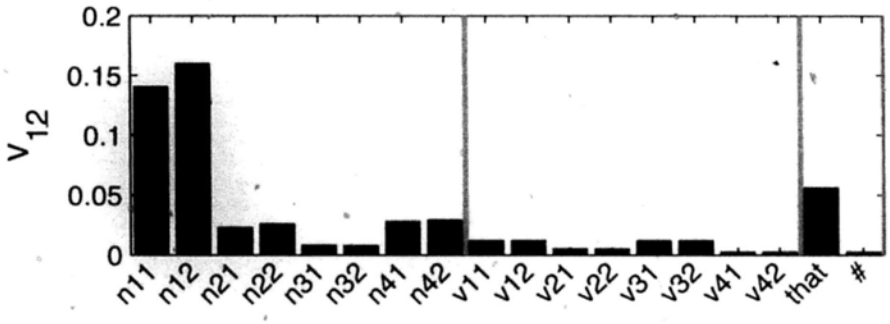



Figure 2.11: Output layer activation at "$n_{31}$-that-$n_{22}$-$v_{41}$-$v_{12}$-..." during the processing of a centre-embedding sentence. The fifth bar chart in penal (b) of Figure 2.8.

words, the networks might have mis-recognised the test sentence as "...-#-$n_{22}$-$v_{41}$-$v_{12}$-...".

### 2.6.4 The first verb of centre-embedding sentences

Perhaps the most challenging sentence position for the artificial networks as well as for humans (Christiansen & Chater, 1999a, and references therein) is the first verb of the centre-embedding sentence. The average output layer activation of the SRNs is shown in the fourth bar chart in penal (b) of Figure 2.8 (reproduced as Figure 2.12, p. 41). The

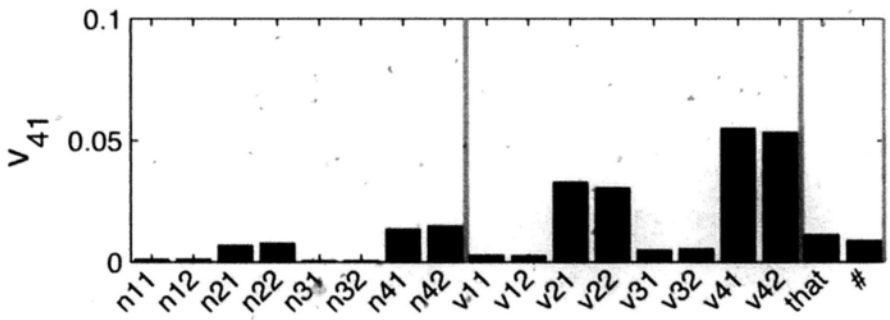



Figure 2.12: Output layer activation at "$n_{31}$-that-$n_{22}$-$v_{41}$-..." during the processing of a centre-embedding sentence. The fourth bar chart in penal (b) of Figure 2.8.

sign of a significant performance degradation can be observed from the overall small

activation values, below 0.06, of the output layer neurons. In spite of the "uncertainty", the output layer activation pattern was different from a bi-gram[16] because verbs were showing much higher activities than nouns.

Another observation of interest as a sign that combinatorial productivity does exist in SRNs, is that at this sentence position, both the group 2 verbs and the group 4 verbs showed relatively high activations. While the transition between a group 4 verb to another group 4 verb was expected since it was in the training set, the transition from a group 4 verb to a group 2 verb was beyond simple bi-gram. Again, the activation of the group 2 verbs reflected the networks memory of the group 2 noun, "$n_{22}$", that wad fed one time step before.

## 2.7 Conclusion and summary of Chapter 2

A computational simulation on *reassessing* the ability of simple recurrent networks (SRNs) to generalise combinatorially was reported. The simulation conducted was inspired by the work of van der Velde and colleagues (2004; 2005) in which they attempted to demonstrate that:

(i) SRNs cannot generalise from training set sentences to sentences that contained new combinations of lexical items.

[16]The second row of Table 2.11 (p. 33)

(ii) SRNs rely on "word-word" association, i.e. the networks rely on bi-gram statistics to process novel sentences, and hence fail to capture the underlying sentence structure.

I consider van der Velde's criticism of SRNs' reliance on bi-gram statistics is worthy of further investigations. Because the essential advancement brought about by SRNs is the incorporation of a recurrent layer that provides the networks ways to accumulate information during the processing of a sequence, in response to the requirements of the task that the networks are trained on (Elman, 1990, 2003). If SRNs do systematically "switch off" the ability to capture sentence it may imply an fundamental flaw of the current connectionist models for sentence processing.

The results obtained from my simulation were contrary to those of van der Velde et al. (2004) that a better performance was attained by the networks. Detailed examination of the output layer activations of the trained networks, which was not done in van der Velde et al. (2004), revealed that bi-gram statistics did not seem to be the dominant information the networks relied on to perform the task of sentence processing. To conclude, the dismissal of SRNs as advocated by van der Velde et al. (2004) was premature but the framework laid out provides new perspectives on the notion of generalisation.

# Chapter 3

# More on SRN Simulation 1

## 3.1 Background

Before I start the discussion on the second simulation in which I aimed to investigate the roles hidden layers in SRNs on the basis of the differences in performance between the networks that I utilised and the ones that van der Velde et al. (2004) utilised, a trivial explanation of the discrepancies needs to be considered, namely the mismatch in the number of trainable weights in the SRNs of my simulation with those in van der Velde et al. (2004).

Recall from Section 2.3.1 (p. 11) that the backbone of an SRN is a layered feed-forward network in which the full connections between layers are trainable with the use of the backpropagation algorithm. It is well known that the performance of the

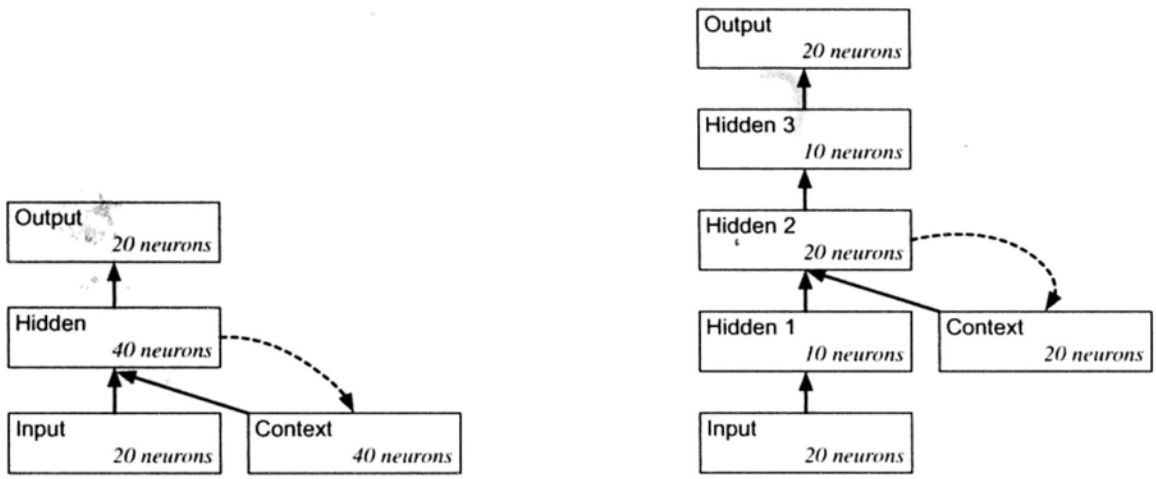



Figure 3.1: The architectures of SRNs used (a) in the simulation reported in Chapter 2 and (b) in van der Velde et al. (2004). Notice that not only two more hidden layers were used in van der Velde et al. (2004), the context layer was coupled with the second hidden layer. Arrows with solid lines denote the trainable connection weights of each network, 3,200 of them in (a) and 1,200 of them in (b).

networks, both in terms of learning the training data and in terms of generalisation, is influenced by the total number of trainable connection weights (Haykin, 1999). Too large a network in size, total number of trainable connection weights, may lead to over fitting and hence generalise poorly; too small a network may result in premature training and hence under-representing the capability of the networks. This is also complicated by the trade-off between the size of the network and the amount of exposure to the training items.

Figure 3.1 (p. 45) shows the architecture of SRNs employed (a) in my simulation reported in Chapter 2 and (b) in the simulation of van der Velde et al. (2004). The two network architectures differed in several aspects:

**Total number of trainable connection weights** For the SRN architecture I used,

there were a total of 3,200 trainable connection weights, 2,400 between the input-context layer and the hidden layer, 800 between the hidden and the output layer. As for the SRN architecture used in van der Velde et al. (2004), there were a total of 1,200 trainable connection weights.

**Number of hidden layers** The architecture I used consisted of a single hidden layer whereas the one van der Velde et al. (2004) used consisted of two more hidden layers.

**The coupling of the context layer with the hidden layer** In both architecture, the context layer was coupled with a hidden layer, namely, the context layer stores the activation of a hidden layer from the previous time step. However, in the architecture employed by van der Velde et al. (2004), the context layer was not coupled with the hidden layer immediately above the input layer, instead it was coupled with the second hidden layer.

In the following section, the above three aspects will be investigated by considering the performance of the networks with various architectures and sizes. These networks were all trained and tested with the procedure described in Chapter 2.

## 3.2 Some more results from SRN-SIM1

### 3.2.1 Concerning the size of the networks

Eighty SRNs with a basic single hidden layer architecture were trained and the GPEs on different types of testing sentences were evaluated and plotted in Table 3.1 (p. 48).

Among the 80 SRNs, the sizes of the networks, i.e. the total number of trainable connection weights, were varied by varying the sizes of the hidden layers. Column (a) of Table 3.1 shows, on the first row, the network configuration of a single hidden layer SRN with 10 hidden layer neurons. On the second row of column (a), the mean GPEs on right-branching sentences attained by twenty such network were plotted and the GPEs on centre-embedding sentences were plotted on the third row. Similarly, GPEs attained by SRNs of different sizes were plotted in second and the third row of column (b, with 20 hidden layer neurons), (c, with 40 hidden layer neurons) and (d, with 80 hidden layer neurons) in Table 3.1.

The figure legend is the same as the one in Figure 2.7 (p. 32) and it is reproduced in Figure 3.2 (p. 49), where black lines denote GPEs on training set sentences; green lines denote GPEs on $M = 2$ testing set sentences; blue lines for $M = 3$; red lines for $M = 4$. For comparison, results reported by van der Velde et al. (2004), GPEs on $M = 3$ testing set sentence, were plotted in grey lines with square marks; and grey lines with circle marks denote the expected GPEs by a bi-gram model. The error bars indicate plus and minus one standard deviation from mean GPEs of the networks I trained.

The GPE plots in column (b), SRNs with 20 hidden layer neurons, of Table 3.1 represents the results of training SRNs that were compatible, in terms of the total number of trainable connection weights, with the ones used in van der Velde et al. (2004). The GPE evaluation of such networks resemble the results reported in Chapter 2 where SRNs with 40 hidden layer neurons, i.e. column (c), were used. This suggested

Table 3.1: GPE evaluation of simple recurrent networks with **one hidden layer** for SRN-SIM1. Networks varied in the total number of trainable connection weights by varying the size of the hidden (context) layer. See Figure 3.2 (p. 49) for the figure legend.

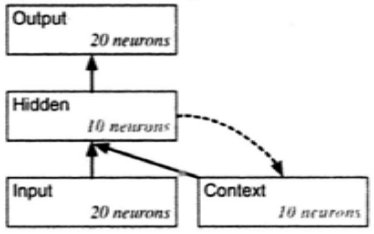
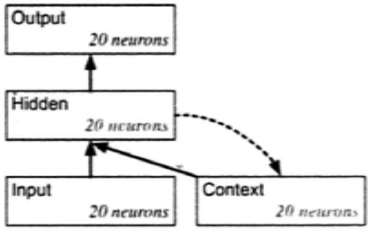
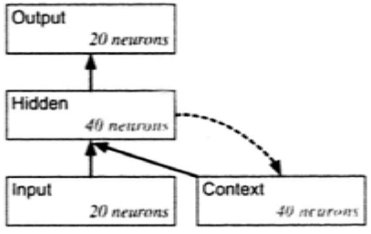
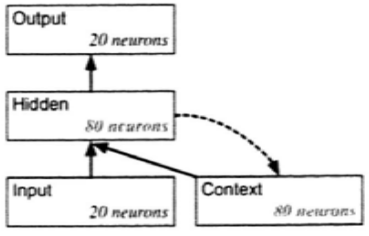
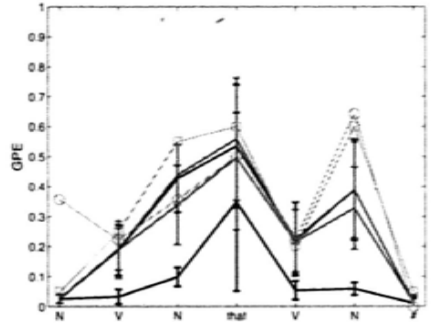
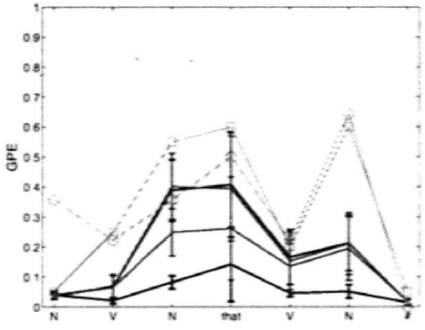
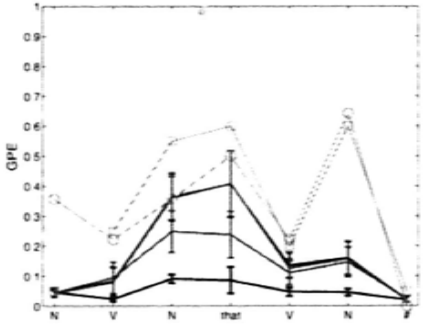
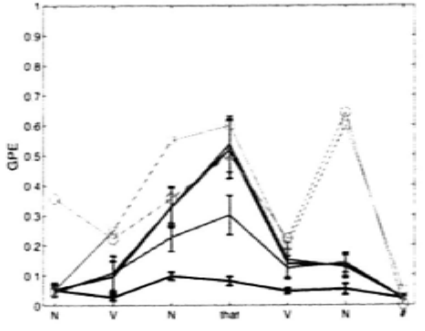
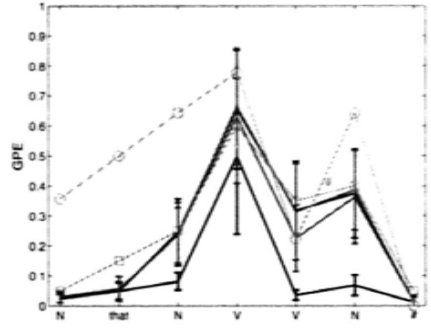
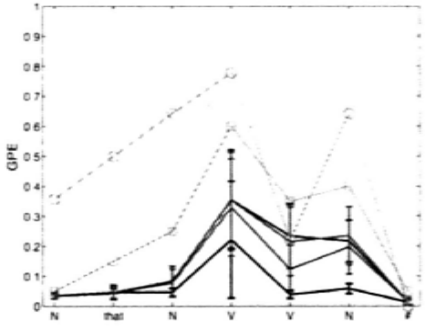
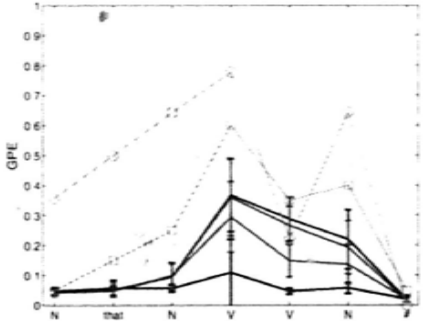
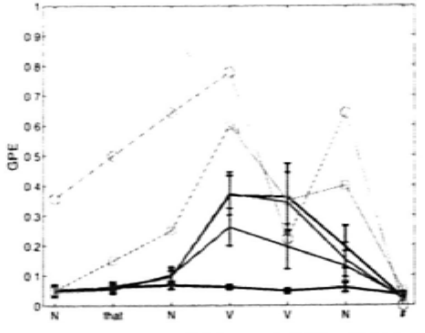
| | **(a)** | **(b)** | **(c)** | **(d)** |
|---|---|---|---|---|
| Network configuration (Total trainable weights) |  (500) |  (1,200) |  (3,200) |  (9,600) |
| Right-branching |  |  |  |  |
| Centre-embedding |  |  |  |  |


Figure 3.2: Legend for Table 3.1

that the discrepancies between my simulation and that of van der Velde et al. (2004) cannot be due to a mismatch in network size.

Although the testing set GPEs of networks with only 500 trainable weights, column (a) of Table 3.1, were almost identical to the GPEs reported in van der Velde et al. (2004), the networks' performance on *training set sentences* was poor at the fourth words of the two complex sentences, which was contrary to both the original study and to the one I reported in Chapter 2. This suggests that these SRNs with only 10 hidden layer neurons had not yet been properly trained. The networks' output layer activations, plotted in Figure 3.3 (p. 50), to a training set sentence provided evidence for that. Because an output layer activation is to be interpreted as probability distributions[1] it should approximately sum to one if the training was successful, at least for the training data. The output layer activations at the fourth sentence position, the fourth row of Figure 3.3 (p. 50), summed to about 0.4 only as oppose to a "successful case" that utilised SRNs with 40 hidden layer neurons, plotted in Figure A.1 (p. 128) in Appendix A.

[1] c.f. Section 2.3.2 (p. 19) on Model training and evaluation

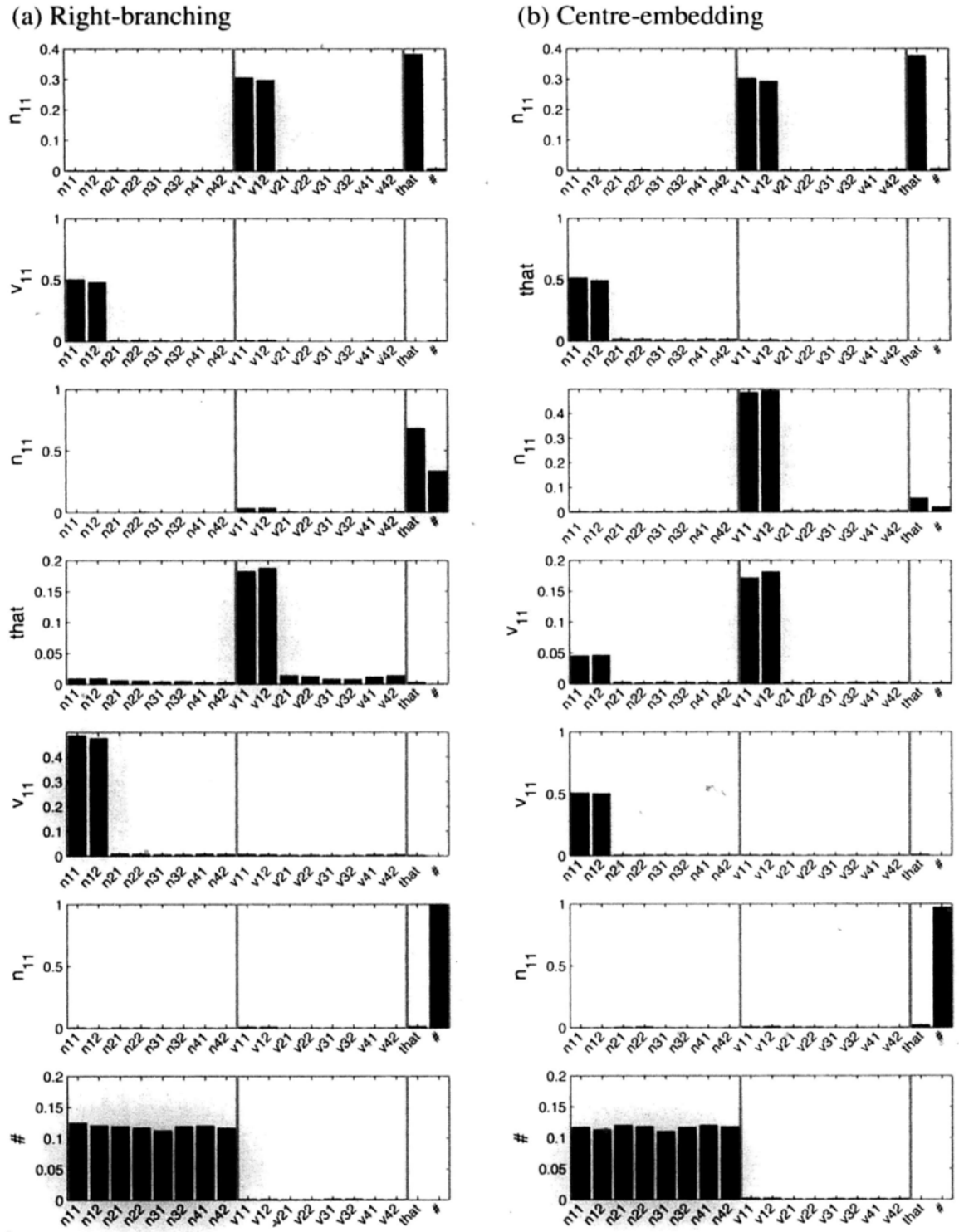



Figure 3.3: The time course of output layer activations of the twenty **single hidden layer** SRNs with 10 hidden layer neurons during the processing of (a) a training set right-branching sentence and (b) a training set centre-embedding sentence. The networks were trained with the procedure described in Chapter 2. Each bar chart represents the output layer activation averaged across the twenty SRNs.

### 3.2.2 Concerning the number of hidden layers and how they were coupled with the context layer

Eighty SRNs with a *two hidden layers* architecture were trained and the GPEs on different types of testing sentences were evaluated and plotted in Table 3.2 (p. 53). Two types of network architecture were used:

**Type I** The context layer was coupled with the hidden layer immediately above the input layer, column (a) and (b) of Table 3.2.

**Type II** The context layer stored the activation of the *second* hidden layer and this activation was fed to the hidden layer immediately above the input layer at the next time step, column (c) and (d) of Table 3.2.

For each network types, twenty SRNs with 4,800 and twenty with 16,000 trainable connection weights were trained and the GPEs on test sentences were averaged across the twenty networks. Similar to Table 3.1 (p. 48), the GPEs evaluation on different types of test sentences for each kind of network were plotted and arranged according to network configurations (columns) and sentence types (rows).

A comparison of the GPE plots of Table 3.2 (a) and (b) with the GPE plots of Table 3.1 (c) and (d) reveals an interesting observation. The GPEs on testing set sentences attained by SRNs with ***two hidden layers were*** smaller than that attained by SRNs with ***a single hidden layer***. SRNs with a second hidden layer seem to perform

better on the testing set sentences and hence showing a better ability to generalise. However, such an improvement was observed only in Type I SRNs where the context layer was coupled with the hidden layer immediately above the input layer. Type II SRNs, column (c) and (d) in Table 3.2, performed worst among the eight network configurations tabulated in Table 3.1 and Table 3.2.

## 3.3 Implications of the analysis

In the last section, I have presented the additional simulation results from SRN-SIM1 with the aim to compare the performance towards combinatorial productivity:

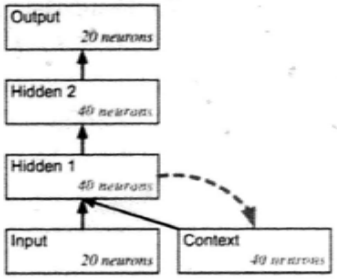
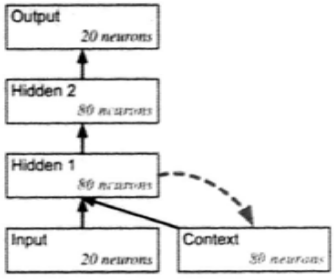
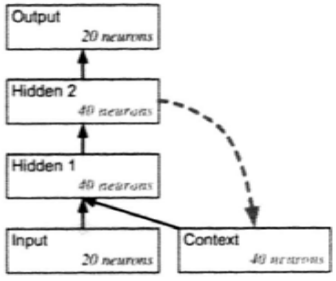
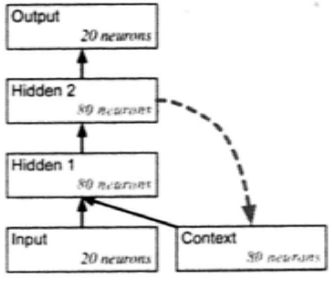
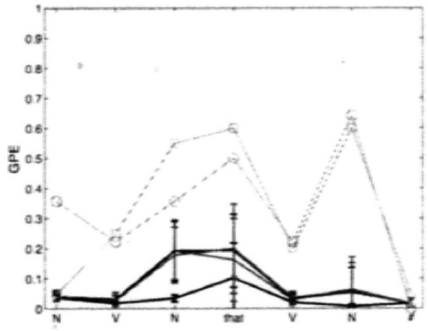
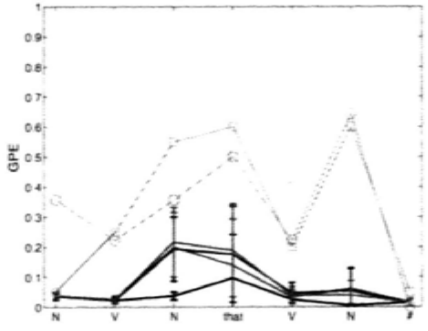
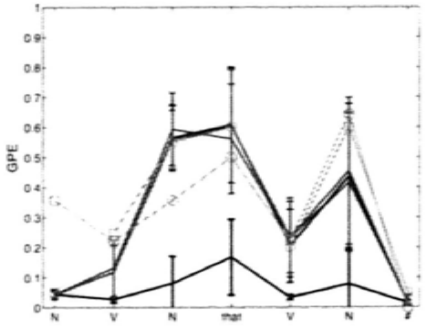
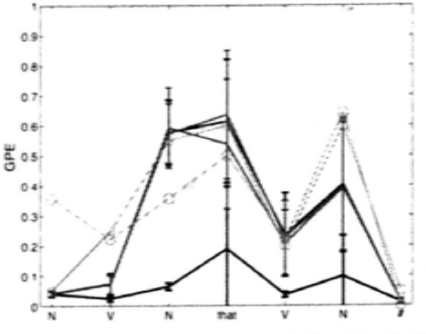
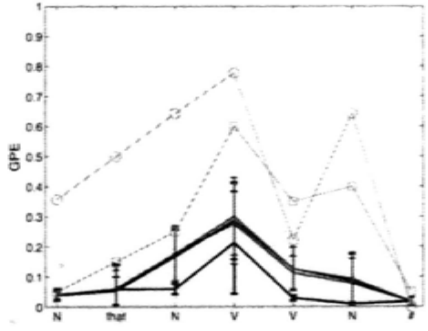
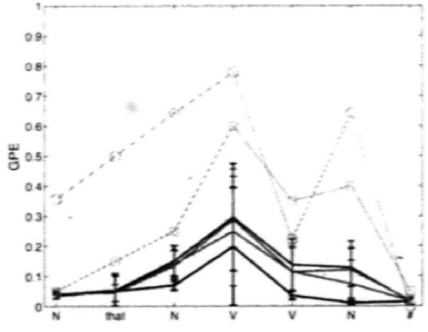
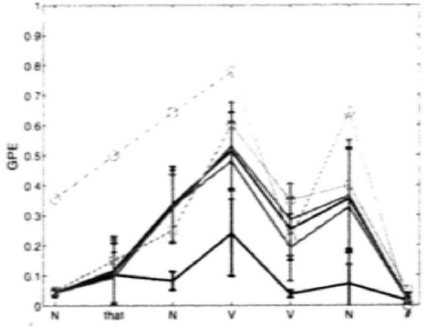
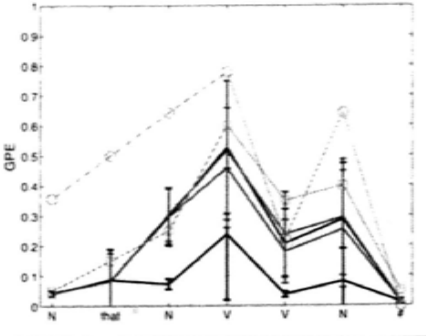
(i) among SRNs with a single hidden layer but differed in the number of hidden layer neurons used and as a result differed in the total number of trainable connection weights

(ii) between SRNs with a single hidden layer and SRNs with a second hidden layer

(iii) between the two ways (Type I and Type II) of coupling of the hidden and context layers

I argue that network size was not the major factor driving the discrepancies between my simulation and that of van der Velde et al. (2004) because the results of the training of SRNs with the same number of trainable connection weights (Table 3.1 (b), p. 48) as used by van der Velde et al. (2004) resulted in GPE evaluations that were qualitatively

Table 3.2: GPE evaluation of simple recurrent networks with **two hidden layers** for SRN-SIM1. Networks varied in the total number of trainable connection weights by varying the size of the hidden (context) layers. See Figure 3.2 (p. 49) for the figure legend.

| | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| Network configuration (Total trainable weights) |   (4, 800) |   (16, 000) |   (4, 800) |   (16, 000) |
| Right-branching |   |   |   |   |
| Centre-embedding |   |   |   |   |

similar to those reported in Chapter 2. By comparing GPE evaluations plotted in Table 3.1 (p. 48) and Table 3.2 (p. 53), I demonstrated that SRNs with a second hidden layer performed better and attained a smaller GPEs, however, this was true only for SRNs in which the context layer was coupled with the hidden layer immediately above the input layer.

# Chapter 4

# Simple Recurrent Network Simulation 2 (SRN-SIM2)

## 4.1 Overview—The tale of two layers in simple recurrent networks

In this chapter I am going to report a computational simulation as an extension of the one reported in Chapter 2.[1] Recall that in Chapter 2, I presented my replication of the study of van der Velde et al. (2004) and reported findings that were contrary to the original study. I argued that SRNs do exhibit combinatorial productivity towards language processing on the basis of their sensitivity to sentence structure.

The simulation was done with identical ways of constructing the training and testing sets sentences as well as the ways to train and evaluate the networks. The

[1] A portion of the findings in the chapter has been reported in the *2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems* (Wong & Wang, 2007b) and in *Dynamics of Continuous, Discrete and Impulsive Systems (Series A): Advances in Neural Networks*, *14*(SI), 650–657. (Wong & Wang, 2007a)

simulation, however, differed from that of van der Velde et al. (2004) in the choice of the architecture of the SRNs, I used SRNs with one hidden layer only, while two more hidden layers were used in van der Velde et al. (2004). This hinted me to conduct this follow-up investigation. In particular, the impacts of utilising multiple hidden layers in SRNs will be investigated and I will conclude that SRNs that were more successful in generalisation developed a "division of labour" among layers. A gross categorisation according to word class was developed in the first hidden layer while a more fine-grained categorisation according to sentence context was developed in the second hidden layer.

## 4.2 Background

The motivation behind this investigation was from an extensive comparison of the networks' performance towards combinatorial productivity among SRNs of different size and of different architectures. This has been discussed in Chapter 3.

## 4.3 Methods

Based on the observation that simple recurrent networks with a second hidden layer and with the context layer coupled with the first hidden layer showed better performance towards generalisation, I conducted the simulation focused on networks with such an optimum configuration.

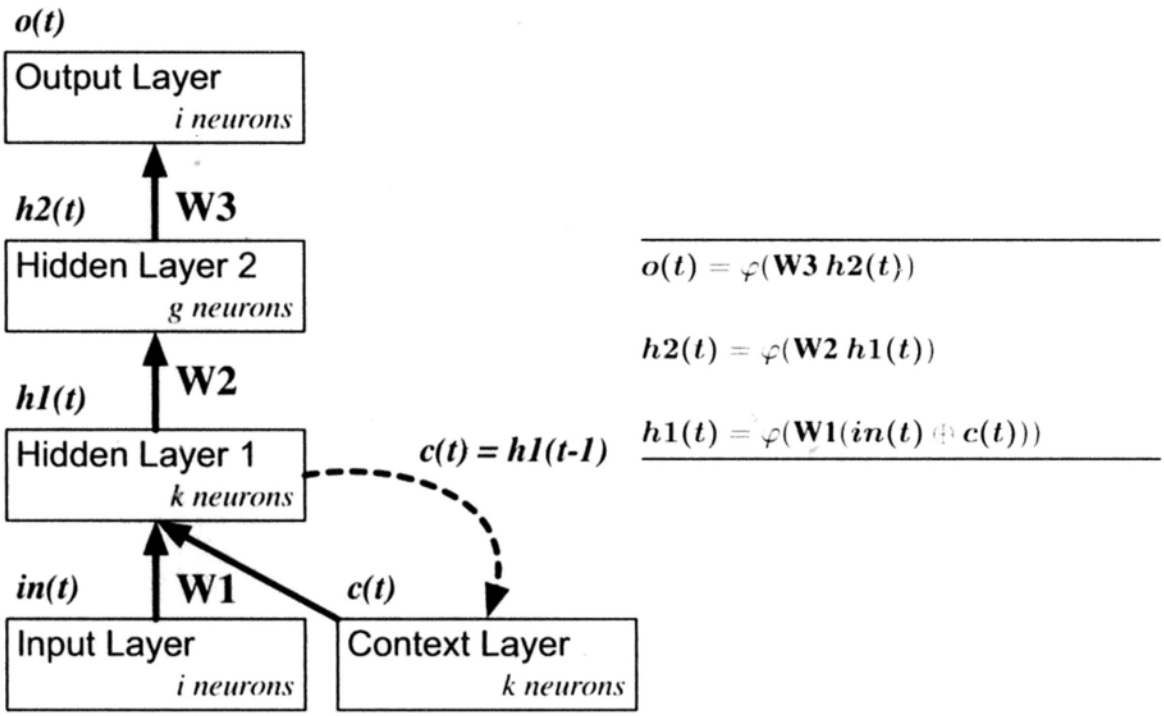



Figure 4.1: The general architecture of a simple recurrent network with two hidden layers (2-hidden-layer-SRNs). The context layer is coupled with hidden layer 1, the hidden layer immediately above the input layer. Arrows with solid lines denote full connection between layers of neurons, represented as blocks. Arrow with dotted line denotes the copy-back one-to-one connections. The directions of signal propagation during the feed-forward operation of an SRN are denoted by the arrows. **W1**, **W2** and **W3**, the three weight matrices, denote the trainable connection weights between the layers of neurons. The equations governing the way signals propagate are given on the right.

Table 4.1: The five-phase training scheme for SRN-SIM2

| | | **Number sentences fed to a network** | | | |
|---|---|---|---|---|---|
| **Phase** | **Token ratio**† | *Simple* | *Right-branching* | *Centre-embedding* | *Total* |
| I | 1 : 0 : 0 | 32,000 | 0 | 0 | 32,000 |
| II | 6 : 1 : 1 | 7,680 | 1,280 | 1,280 | 10,240 |
| III | 2 : 1 : 1 | 25,600 | 12,800 | 12,800 | 51,200 |
| IV | 1 : 2 : 2 | 12,800 | 25,600 | 25,600 | 64,000 |
| V | 1 : 2 : 2 | 128,000 | 256,000 | 256,000 | 640,000 |
| **Total** | | 206,080 | 295,680 | 295,680 | 797,440 |

†Ratio of simple : right-branching : centre-embedding sentences

Figure 4.1 (p. 57) shows the general two hidden layer architecture employed in this simulation. SRNs of this kind are denoted as *2-hidden-layer-SRNs* hereafter.

I extended the four-phase training scheme (Table 2.9, p. 28) into a five-phase one. Table 4.1 (p. 58) lists the number of sentences fed to the networks for each of the phase. Phase I to Phase IV were the same as they were in SRN-SIM1. Phase V was an extension of Phase IV by repeating it for ten times. The total number sentences fed to the networks was increased from 157,440 to 797,440. This four-fold increment aimed to compensate for the increase in network size due to the additional hidden layer and to exclude the possibility of premature training that might lead to an underestimate of networks' performance.[2]

Four network configurations were evaluated in this simulation and they were listed

[2]It should be noted that there was no attempt to approximate the amount of input data a child receives during the process of actual language acquisition. The "amount of training" a network receives by feeding a sentence to a network once is difficult, if at all possible, to translate into an "amount of learning" a child receives by listening to a sentence once. The aim of the simulation was to investigate if SRNs are computationally adequate for generalisation.

Table 4.2: Summary of network configurations evaluated in SRN-SIM2, using SRN with two hidden layers

| | **Number of neurons in hidden layer** | | |
|---|---|---|---|
| **Network configuration** | *First* | *Second* | **Total trainable weights** |
| (i) | 40 | 20 | 3,600 |
| (ii) | 40 | 20 | 4,800 |
| (iii) | 80 | 40 | 12,000 |
| (iv) | 100 | 40 | 16,800 |

in the first row of Table 4.3 (p. 61), briefly summarised in Table 4.2 (p. 59).

For each network configuration, twenty networks were constructed and each was initialised with an independent random set of connection weights with values drawn from a normal distribution with a mean of zero and a standard deviation of 0.05. The learning rate and momentum was set to 0.1 and 0, respectively.

These eighty 2-hidden-layer-SRNs were trained and evaluated using the procedure described in Chapter 2. The networks were trained with the prediction task with streams of concatenated sentences which were randomly sampled from the training sets. The results reported in the remaining parts of the chapter were based on the performance averaged across twenty networks for each network configuration.

## 4.4 Results

After the fifth phase of training, the connection weights of the networks were frozen. Testing set sentences of various types and M-values[3] were fed to the networks through which the output activations of the networks were subject to GPE evaluation on grammaticality.

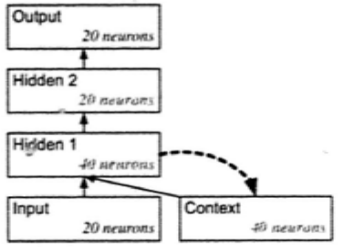
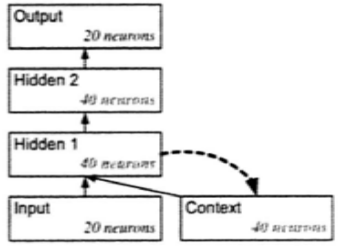
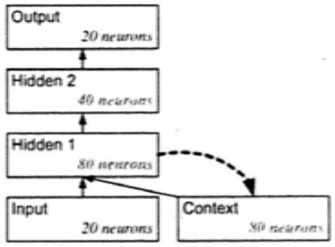
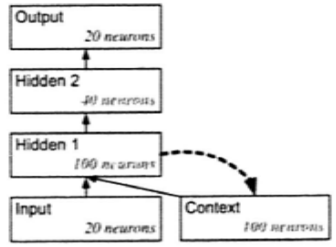
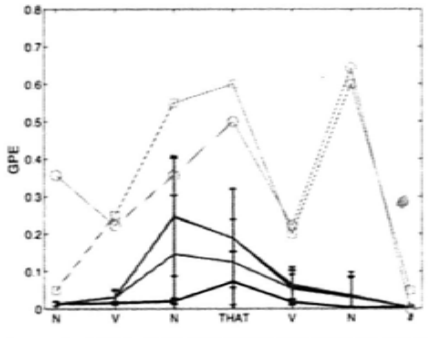
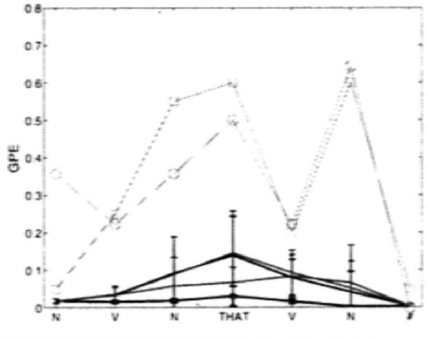
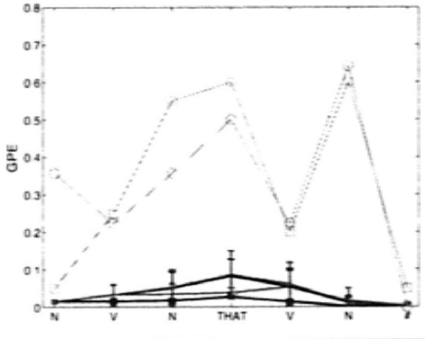
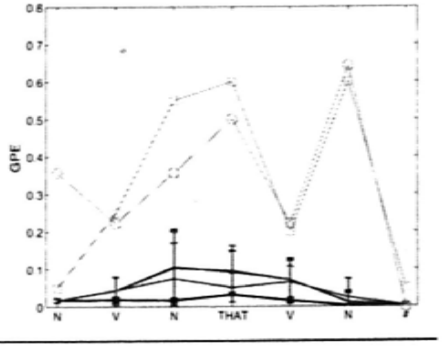
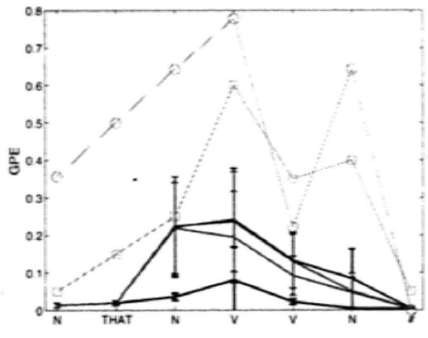
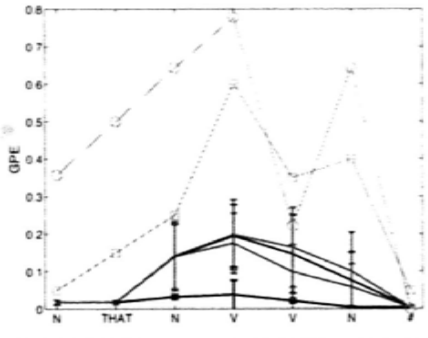
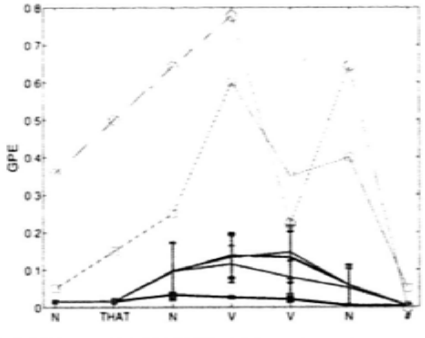
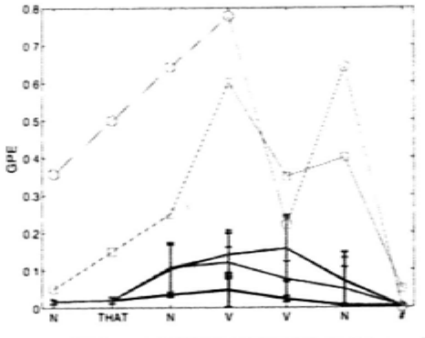
Each network was evaluated with 100 sentences drawn randomly from each subset, sentence type by M-value,[4] of the testing set sentences. The mean GPEs attained by each SRN at each sentence position were recorded. For each network configuration, the grand mean GPEs were obtained by taking the average across the twenty networks and they were plotted in Table 4.3 (p. 61).

Except for the lines in grey colour, a data point in the figure represents the GPEs of SRNs' outputs in predicting the next lexical item at the sentence position indicated on the $x$-axis. For example, the first data point of each graph in Table 4.3 (p. 61) represents the grand mean GPE of networks' outputs given the sentence-initial nouns were fed as the inputs and at this sentence position either a verb or the relative marker "`that`" was considered to be the grammatically correct prediction. The figure legend for Table 4.3 is given in Figure 4.2 (p. 62).

Compare with the GPE evaluation reported in SRN-SIM1 in Chapter 2,[5] the 2-hidden-layer-SRNs employed in this simulation showed a great deal of improvement.

[3] $M$ is the number of groups of lexical items that were mixed to construct the testing sets
[4] cf. The construction of testing sets in Table 2.10 (p. 30)
[5] cf. Figure 2.7 (p. 32) and Table 3.1 (p. 48)

Table 4.3: GPE evaluation of simple recurrent networks with **two hidden layers** for SRN-SIM2. Networks varied in the total number of trainable connection weights by varying the size of the hidden (context) layers. See Figure 4.2 (p. 62) for the figure legend.

| | (i) | (ii) | (iii) | (iv) |
|---|---|---|---|---|
| Network configuration (Total trainable weights) |  (3, 600) |  (4, 800) |  (12, 000) |  (16, 800) |
| Right-branching |  |  |  |  |
| Centre-embedding |  |  |  |  |

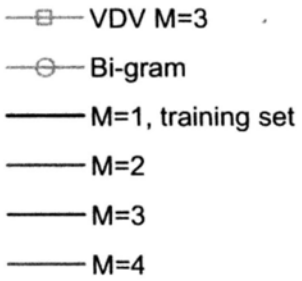



Figure 4.2: Legend for Table 4.3 and Table 4.4

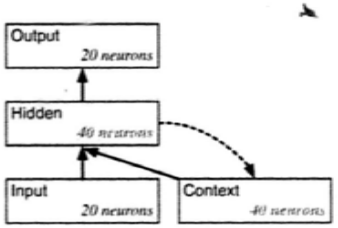
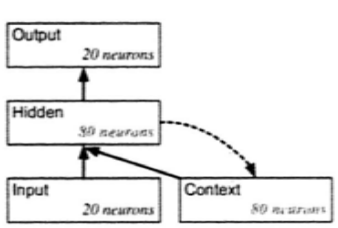
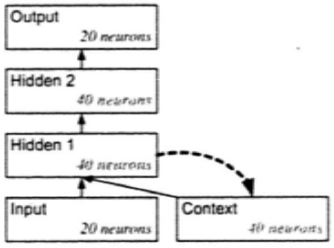
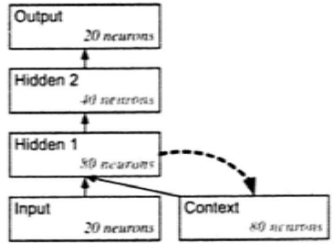
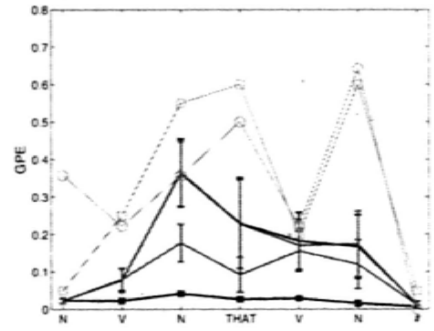
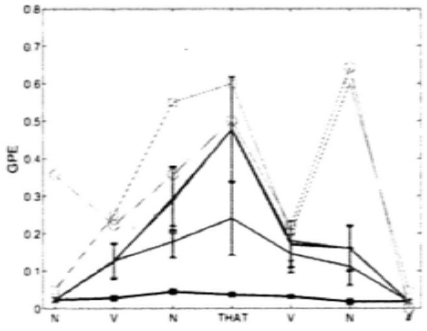
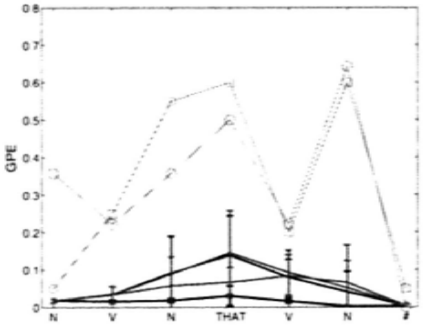
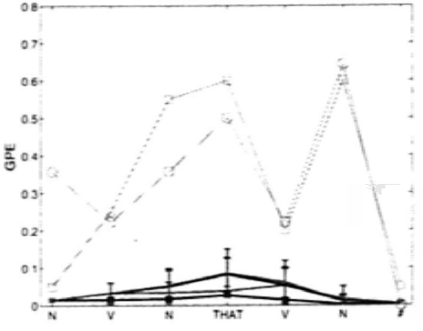
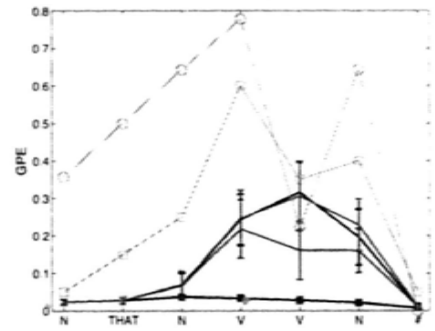
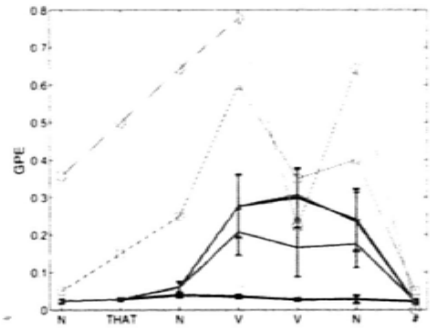
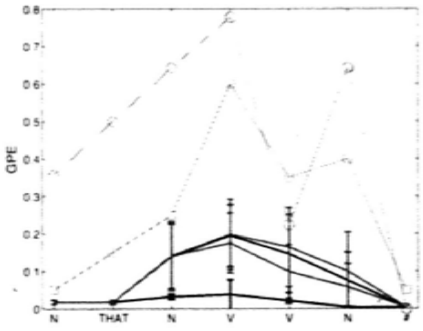
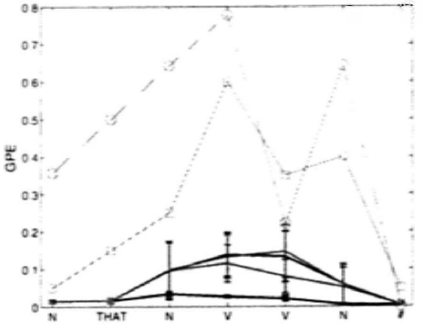
For networks beyond a certain critical mass, network (ii), (iii) and (iv) in Table 4.3, the mean GPEs at all sentence positions were all below 0.2. Although centre-embedding sentences remained to be difficult to process which was consistent with human performance (Gibson & Pearlmutter, 1998; Hsiao & Gibson, 2003; Gibson, 2000).

The significant improvement over the results reported by van der Velde et al. (2004), grey line with square markers, provided a strong evidence against the notion of the "*lack of combinatorial productivity*" as advocated by van der Velde et al. (2004). At the very least, the claim of van der Velde et al. (2004) could not be generalised to all SRNs neither could it be generalised to all connectionist models.

### 4.4.1 Comparison with SRNs with a single hidden layer

To ensure an unbiased comparison with SRNs with a single hidden layer, forty *single-hidden-layer-SRNs* were trained with the revised training scheme. Twenty of them with 40 hidden layer neurons and twenty of them with 80 hidden layer neurons. The GPE evaluations of these networks were plotted in column (a) and (b) of Table 4.4 (p. 63).

Table 4.4: GPE evaluation of simple recurrent networks with (a) and (b) **a single hidden layer** and (c) and (d) **two hidden layers** for SRN-SIM2. Networks varied in the total number of trainable connection weights by varying the size of the hidden (context) layers. See Figure 4.2 (p. 62) for the figure legend.

| (a) | (b) | (c) | (d) |
|---|---|---|---|
| Network configuration (Total trainable weights) | | | |
|  |  |  |  |
| (3, 200) | (9, 600) | (4, 800) | (12, 000) |
| Right-branching | | | |
|  |  |  |  |
| Centre-embedding | | | |
|  |  |  |  |

GPE evaluations of 2-hidden-layer-SRNs of comparable sizes were reproduced in column (c) and (d) of Table 4.4 to ease the comparison.

The statistical analyses in the next section were based on the set of data graphed in Table 4.4 which involved network configurations of four kinds, summarised in Table 4.5 (p. 64). Two of which were SRNs with a single hidden layer and two of which were SRNs with two hidden layers.

Table 4.5: Summary of network configurations evaluated in SRN-SIM2, for a comparison between SRNs with a single hidden layer (a) and (b) and SRNs with two hidden layers (c) and (d). cf. Table 4.4 (p. 63)

| | **Number of neurons in hidden layer** | | |
|---|---|---|---|
| **Network configuration** | *First* | *Second* | **Total trainable weights** |
| (a) | 40 | N/A | 3,200 |
| (b) | 80 | N/A | 9,600 |
| (c) | 40 | 40 | 4,800 |
| (d) | 80 | 40 | 12,800 |

The overall mean GPEs, collapsed across sentence types, achieved by each of these 80 networks to training set sentences and $M = 4$ testing set sentences were also plotted in Table 4.6 (p. 65) and Table 4.7 (p. 66) for SRNs with a single hidden layer and with two hidden layers, respectively.

## 4.5 Analysis of GPE

Interestingly, although both kind of networks showed a sign of near perfect learning on training set sentences, all networks achieved a less than 0.04 overall mean GPE,

Table 4.6: Overall mean GPEs on training set and $M = 4$ testing set sentences attained by each of the twenty SRNs with network configuration (a) and (b). The height of the upper portion of an error bar denotes one standard deviation.

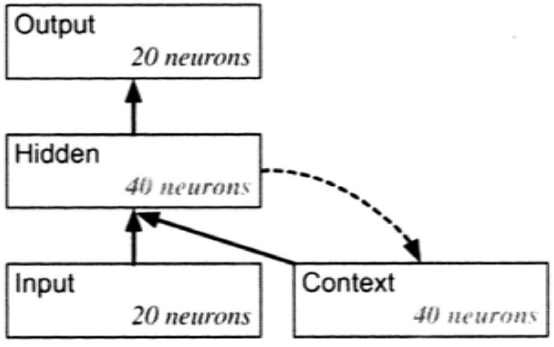
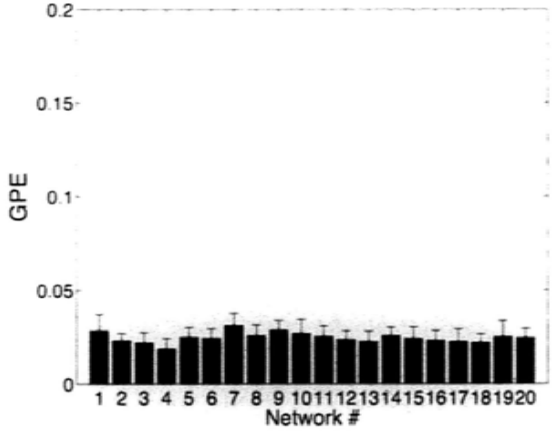
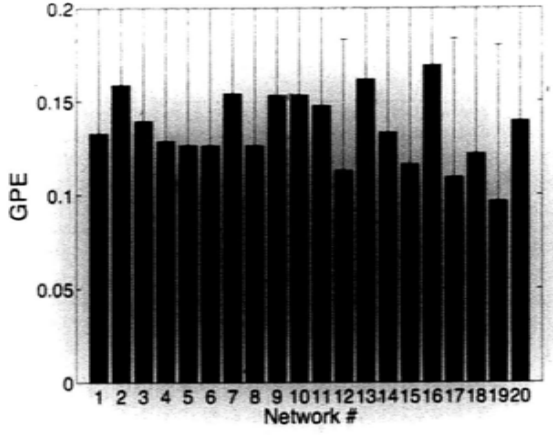
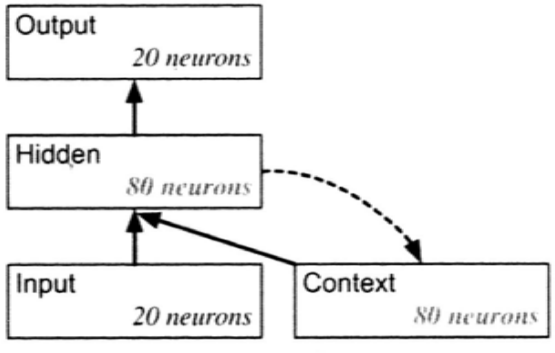
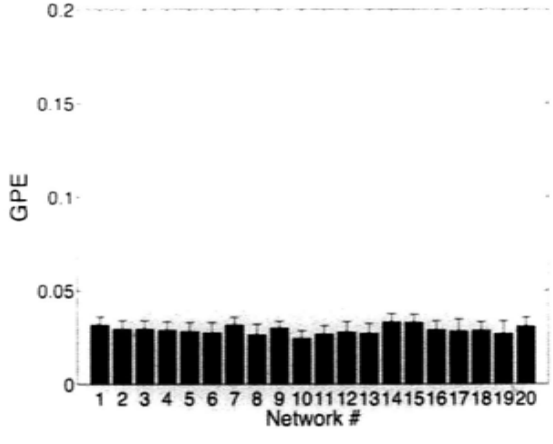
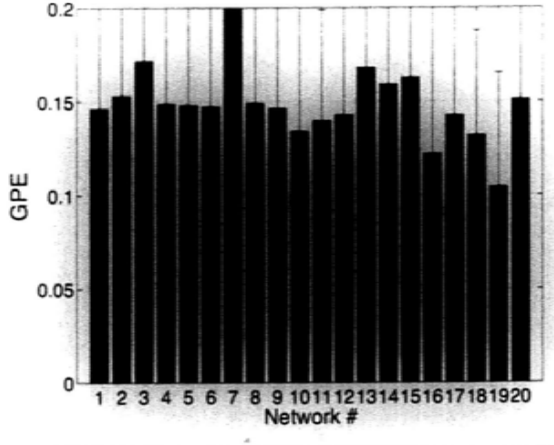
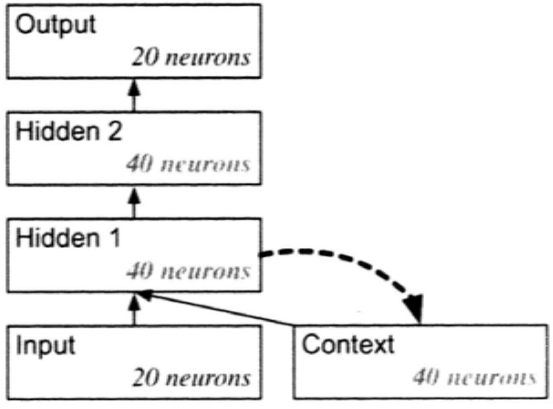
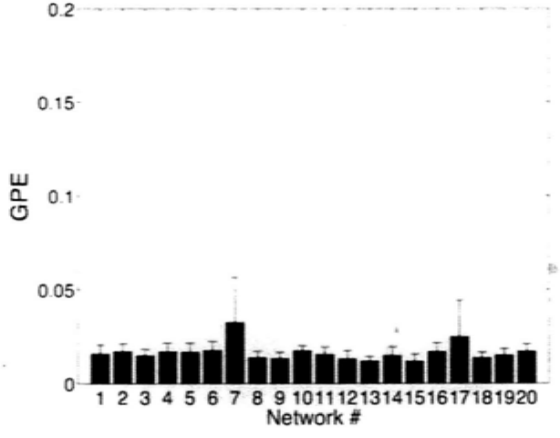
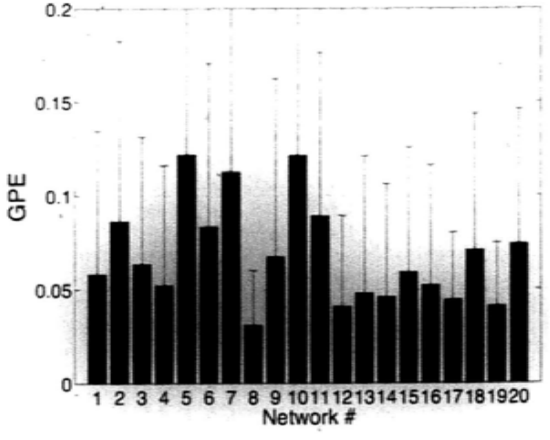
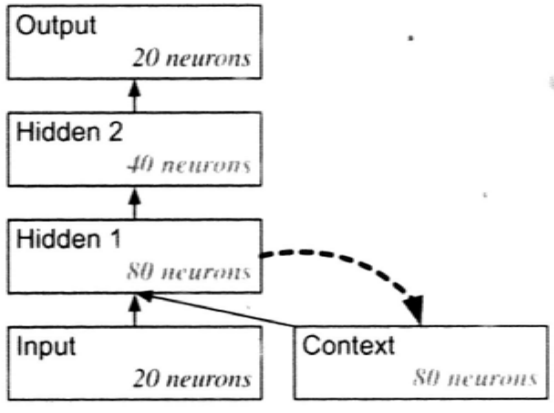
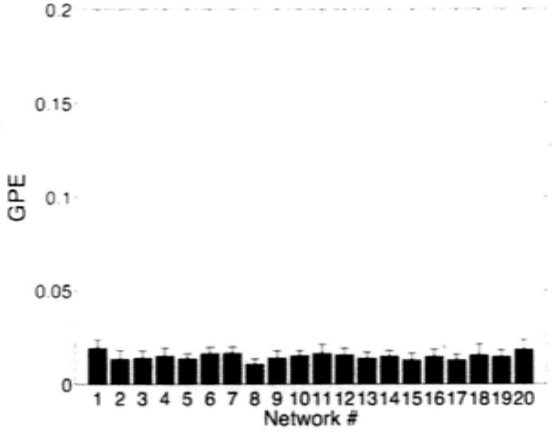
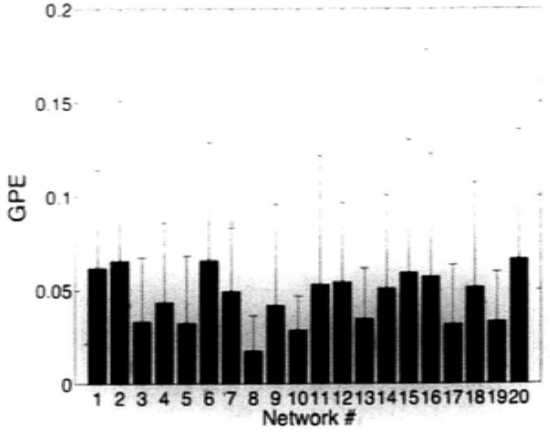
| **Network Configuration** | **Training set sentences** | **Testing set sentences,** $M = 4$ |
|---|---|---|
| As in Table 4.4 (a)   |   |   |
| As in Table 4.4 (b)   |   |   |

Table 4.7: Overall mean GPEs on training set and $M = 4$ testing set sentences attained by each of the twenty SRNs with network configuration (c) and (d). The height of the upper portion of an error bar denotes one standard deviation.

| **Network Configuration** | **Training set sentences** | **Testing set sentences**, $M = 4$ |
|---|---|---|
| As in Table 4.4 (c)   |   |   |
| As in Table 4.4 (d)   |   |   |

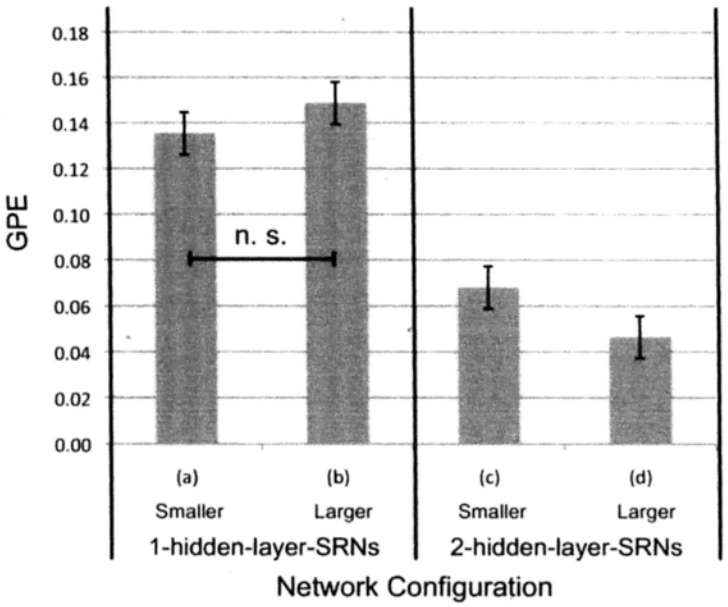
networks with two hidden layers showed considerable smaller GPEs in processing testing set sentences. This observation was supported by a 2-by-2 analysis of variance (ANOVA), with LAYER (1-hidden-layer-SRNs, 2-hidden-layer-SRNs) and SIZE (smaller, larger) as the two factors, on the overall mean GPE of each of the networks in Table 4.6 and Table 4.7 as the dependent variable.

The main effect of LAYER was significant, $F(1, 76) = 333.96$, $p < 0.001$, indicating that 2-hidden-layer-SRNs attained a smaller GPE than did 1-hidden-layer-SRNs. The LAYER $\times$ SIZE interaction effect was also significant, $F(1, 76) = 14.22, p < 0.001$ and a post-hoc analysis suggested that 2-hidden-layer-SRNs of larger size attained a smaller GPE than did 2-hidden-layer-SRNs of smaller size, but there was no such trend for SRNs with a single hidden layer.

The post-hoc analysis was conducted as an one-way ANOVA with network configuration (Table 4.4 (a), (b), (c) and (d)) as the fixed factor and the result of the analysis was plotted in Figure 4.3 (p. 68), in which the estimated marginal means were given together with the 95% confidence intervals marked by the error bars. All of the pair-wise comparisons were significant[6] except between network configuration (a) and (b).

To summarise, networks with only one hidden layer showed no sign of improvement in response to the extended training, as compared to SRN-SIM1. The difference in

[6]Adjustment for multiple comparisons: Bonferroni




| N. C.† | Mean | S. D.‡ |
|---|---|---|
| (a) | 0.135 | 0.019 |
| (b) | 0.149 | 0.020 |
| (c) | 0.068 | 0.027 |
| (d) | 0.047 | 0.014 |

† Network configuration
‡ Standard deviation

Figure 4.3: Estimated marginal mean GPEs achieved by SRNs of four network configurations. cf. Table 4.4 (p. 63). Error bars indicate 95% confidence intervals obtained by one-way ANOVA as the post-hoc analysis. The table on the right gives the values and the standard deviations. n. s.: $p > 0.05$, see text for details.

performance between 1-hidden-layer-SRNs and 2-hidden-layer-SRNs was significant.[7]

What could be more informative is the relatively large variation in performance on testing set sentences among SRNs of network configuration (c). The GPEs attained by these 2-hidden-layer-SRNs ranged from as small as 0.031 to as large as 0.122, the first row of Table 4.7 (p. 66). Such an individual difference among networks of the same type together with the difference in performance between SRNs with two hidden layers and SRNs with a single hidden layer provides an opportunity to explore if there exist a qualitative difference between successful and less successful networks which may shed

[7]When comparing the performance of the networks, I did not aim to control for the effect of over-training and under-training. However, a post hoc analysis on network performance suggested that the main conclusion drawn in this chapter was not affected by the choice of the stopping criterion.

Consider the contrast between network configurations (a) and (c) in Table 4.4 (p. 63) for example, network configuration (a) did not show signs of over-training since the performance on testing set GPE was no worse than that in SRM-SIM1 (Table 3.1 (p. 48)), where the networks were training with less tokens. For all network configurations with results tabulated in Table 4.4 (p. 63), there were no sign of under-training either as the mean training set GPEs were close to zero with very small variances also.

light on the plausible mechanism that underlines generalisation. This is to be discussed in the next section.

## 4.6 Combinatorial productivity through the emergence of categories

Recall that the essence of combinatorial productivity is *generalisation*. This kind of generalisation is argued to be essential for acquiring human language because of the *combinatorially productive* nature of natural language. Such kind of productivity arises in a way that the meaning space grows multiplicatively with the size of the lexicon and the syntactic complexity.[8] In the modelling framework outlined in previous chapters, combinatorial productivity is about extending the process of extracting the predicate-agent-patient meaning from a few exemplars to all possible combinations of lexical items in order to keep up with the ever growing meaning space.

Generalisation undoubtedly does not come out of the blue, it depends on similarities between instances in the training set and instances in the testing sets. In light of this, proponents of a rule-based account of language acquisition (van der Velde & de Kamps, 2006; van der Velde et al., 2004; Marcus et al., 1999) have rightly questioned if connectionist models, SRNs being one such, can exhibit such an ability to generalise.

In the context of the current study, a testing set sentence such as "$n_{1*}$-$v_{2*}$-$n_{3*}$"[9] is

[8] cf. Section 2.2.1 (p. 8). Combinatorial complexity in terms of number.

[9] cf. Section 2.4.1 (p. 24). The lexicon of nouns and verbs was divided into four non-overlapping groups, "$n_{ij}$" denotes the $j^{\text{th}}$ member of the $i^{\text{th}}$ group of nouns and similarly "$v_{ij}$" for verbs. In addition, "$n_{i*}$" and "$v_{i*}$" denotes any $i^{\text{th}}$ group nouns and verbs, respectively.

novel to the networks since the noun "$n_{3*}$" has never been seen by the networks as an object in a sentence in which "$n_{1*}$" is the subject. Recall that the networks were trained with sentences composed by lexical items from the same group, i.e. the networks have seen "$n_{1*}$-$v_{1*}$-$n_{1*}$" and "$n_{3*}$-$v_{3*}$-$n_{3*}$" but not "$n_{1*}$-$v_{2*}$-$n_{3*}$". Therefore the scenario that van der Velde et al. (2004) might argue for was that in processing the testing set sentence, to the networks, the word "$n_{3*}$" appears in an unknown context, "$n_{1*}$-$v_{2*}$-...", which bears no similarity to any of the partial training set sentences and therefore the networks fail to generalise. van der Velde et al. (2004) argued that the remaining cues that the network could make use of are the bi-gram statistics, which, according to the scenario, is true since a basic feed-forward architecture is the major building block of a simple recurrent network (cf. Figure 2.2 in p. 12).

As I have mentioned, generalisation is about making use of what is common between the training space and the testing space. This is also referred to as finding an overlap in representations (McClelland & Plaut, 1999). The anticipation of van der Velde et al. (2004) that SRNs cannot exhibit combinatorial productivity was based on the assumption that the required overlapping for generalisation was absent in the training data.

Despite the attempt of van der Velde et al. (2004) in providing simulation results to support their argument, my simulation results that utilised the same paradigm in evaluating how well SRNs generalise provided an existence proof that SRNs *can* generalise. The question now becomes "how?".

I do not argue that SRNs always generalise to a degree that matches human performance for my simulation results summarised in Table 4.3[10] and Table 4.4[11] demonstrated that the degree of success in generalisation varies. The variation was larger between SRNs with a single hidden layer and SRNs with two hidden layers, the latter showed greater ability to generalise. However, networks with the same configuration also varied in testing set GPEs they attained.

This variability provides a chance to investigate what additional computation successful networks had performed that less successful networks were less able to do so. Grounded on established researches that SRNs are capable of extracting distributional statistics to support computations that require them to keep a memory trace in processing sequences (Grüning, 2006; Rodriguez, 2001; Rodriguez, Wiles & Elman, 1999; Redington et al., 1998; Redington & Chater, 1998b), I speculate that in addition to forming the categories of nouns and verbs, as demonstrated by Elman (1990), categorisation according to sentence positions may also be the driving force for the success of the networks to generalise combinatorially. To be more specific, the success of the network to generalise might depend on the ability to induce that "$n_{31}$" in a testing set sentence is similar to a "$n_{31}$" in a training set sentence in a sense that this word is in an object position regardless of the *semantic context*, i.e. regardless of exactly what words "$n_{31}$" is preceded by. This sort of similarity was implicit in the training data and required the networks to actively induce it.

[10] in p. 61, where performance on generalisation were contrasted among SRNs with two hidden layers.
[11] in p. 63, where performance on generalisation were contrasted between SRNs with a single hidden layer and SRNs with two hidden layers.

Proponents of connectionist models for language acquisition (Elman, 2003; McClelland & Plaut, 1999) have long been arguing that networks are more than passive statistics gathering devise, which the rule-based school often assums. Attempts have been made to show that networks could go beyond surface similarity towards successful generalisation. Elman's early attempt (Elman, 1990) 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 also advocated in McClelland & Plaut (1999):

> *"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."*
>
> McClelland & Plaut (1999, p. 167)

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 $\boldsymbol{h}(t)$ in Section 2.3.1 in p. 11), developed by the networks through training. Since hidden layer activations are of high dimension, method of dimensionality reduction such as principal components analysis (PCA), and hierarchical clustering analysis are often used.

In the remaining part of this chapter, I will contrast the hidden layer activations developed by networks that were more successful in generalisation with those

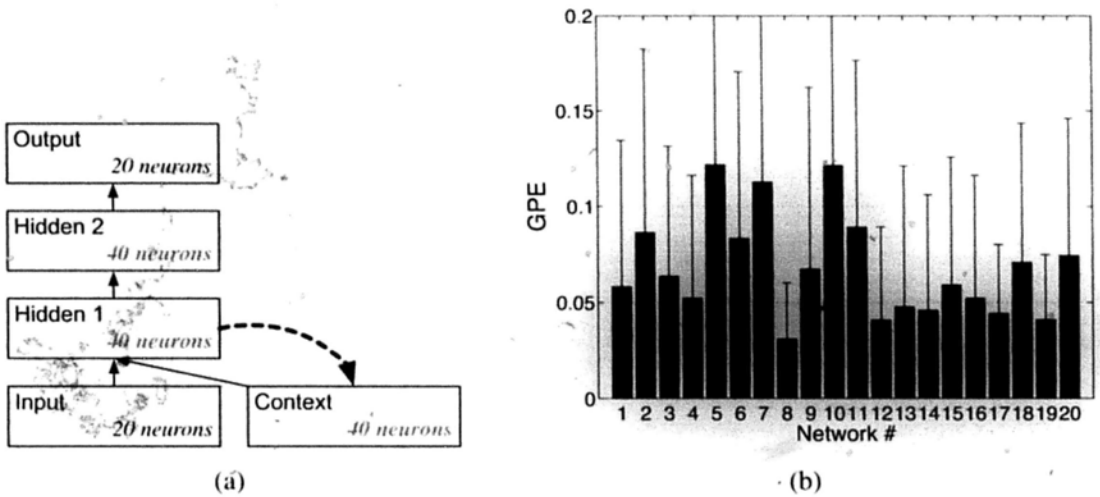



Figure 4.4: Overall mean GPEs (b) on $M = 4$ testing set sentences attained by each of the twenty SRNs with network configuration as shown in (a). The height of the upper portion of an error bar denotes one standard deviation.

developed by networks that were less so. I will focus on the SRNs with the network configuration as shown in Figure 4.4 (a) (p. 73) among which the variation in GPE they achieved was the largest.[12]

## 4.7 Analysis of hidden layer activations

### 4.7.1 Methods

For each sentence type,[13] twenty sentences were randomly sampled from each of the four sets: the training set, $M = 2$, $M = 3$ and $M = 4$[14] testing sets. These 220 sentences were fed to each of the networks trained with the five-phase training scheme

[12] cf. Figure 4.3 (p. 68)

[13] Simple, right-branching and centre-embedding sentences

[14] Except for simple sentences as there were only three lexical items to compose a simple sentence

described in this chapter. For each sentence, the hidden layer activations when a give word was fed to a network as the current word were recorded.

In processing a sentence with a length of $l$, for example, at the first time step, $t = 1$, the first word was fed to the SRN and as signals propagated from the input-context layer to the first hidden layer, the activations of the first hidden layer, $\boldsymbol{h1(t)}$, was obtained.[15] The activations of the second hidden layer, $\boldsymbol{h2(t)}$, was obtained as signals propagated from the first hidden layer to the second hidden layer. Hidden layer activations for the other words in the sentence, $\boldsymbol{h1(t)}$ and $\boldsymbol{h2(t)}$ for $t = 2, ..., l$, were computed in a similar way.

For each network, two matrices **H1** and **H2** were obtained after feeding the set of 220 sentences to the SRN, where each row of the matrices was the hidden layer activation, of the corresponding layer, when a particular word was fed during the testing phase. **H1** and **H2** were of the dimension $1,360$ rows[16] by $40$ columns[17].

### 4.7.2 Results

Principal components analysis (PCA) was applied independently to **H1** and **H2** obtained from each network to reduce the dimensionality from $1,360$-by-$40$ to $1,360$-by-two. Hidden layer activations obtained from SRN #8, which showed the best

[15] cf. Figure 4.1 (p. 57), the general architecture of a 2-hidden-layer-SRN.

[16] Number of test sentences × the length of each sentence, including the end-of-sentence markers. For simple sentences, there were $60 \times 4 = 240$ rows; For right-branching sentences, there were $80 \times 7 = 560$ rows; For centre-embedding sentences, there were also $80 \times 7 = 560$ rows. Altogether there were $240 + 560 + 560 = 1,360$ rows of hidden layer activations for each hidden layer.

[17] Since the first and the second hidden layer contained $40$ neurons

generalisation ability, achieving the lowest overall testing set GPE of 0.031 among the networks of same network configuration,[18] were plotted as grey dots on a scatter plot in penal (b) and (d) of Table 4.8 (p. 76) and Table 4.9 (p. 77). Whereas the hidden layer activations obtained from the SRN that was least successful in generalisation, SRN #5, were plotted in penal (a) and (c) of Table 4.8 (p. 76) and Table 4.9 (p. 77).

Highlighted on the scatter plots in Table 4.8 (p. 76), with coloured symbols shown in the legend, were the hidden layer activations corresponding to the processing of the twenty $M = 4$ right-branching test sentences. The first word of a right-branching sentence was denoted N1 and the second word was denoted V2, the remaining words were denoted in a similar fashion where the letter denoted the word class, N for nouns, V for verbs and T for the relative marker "`that`", and the number denoted the sentence position. The GPE evaluations of the two networks were given in the plots at the bottom row of the table right to the figure legend.

Likewise, highlighted on the scatter plots in Table 4.9 (p. 77), with coloured symbols shown in the legend, were the hidden layer activations corresponding to the processing of the twenty $M = 4$ centre-embedding test sentences.

In all of the scatter plots, a grey rectangle was centred at the centroid of a corresponding highlighted data points grouped according to sentence position. The width and height of a rectangle spanned two standard deviations in PCA scores of the first, PCA(1), and second, PCA(2), principal component, respectively.

[18] cf. Figure 4.4 (p. 73), overall mean GPEs on $M = 4$ testing set sentences attained by each of the twenty SRNs.

Table 4.8: Hidden layer activations during the processing of $M = 4$ **right-branching sentences** obtained from a 2-hidden-layer-SRN that was least successful in generalisation (network #5) and a 2-hidden-layer-SRN that was most successful in generalisation (network #8), see text for details.

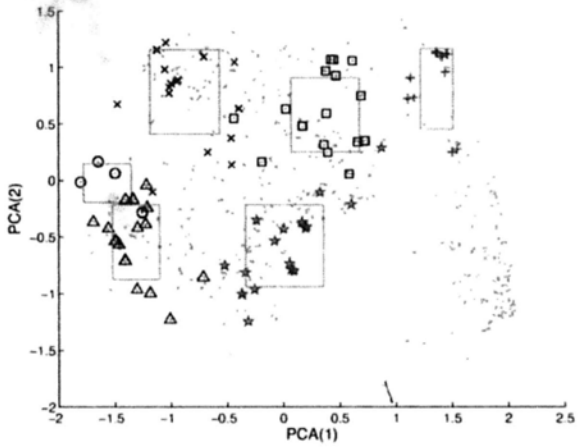
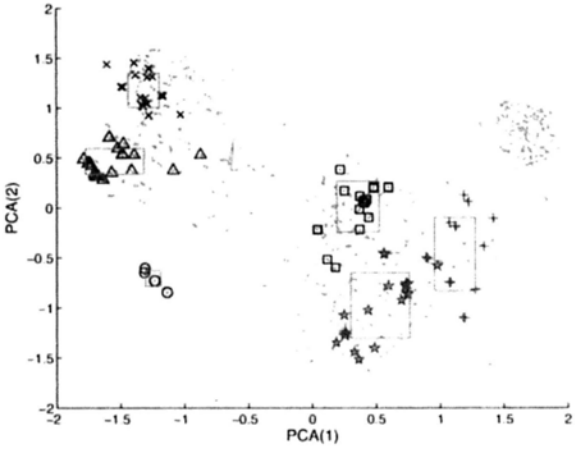
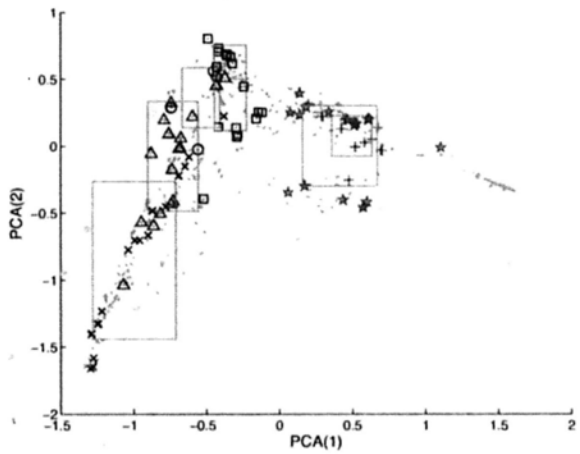
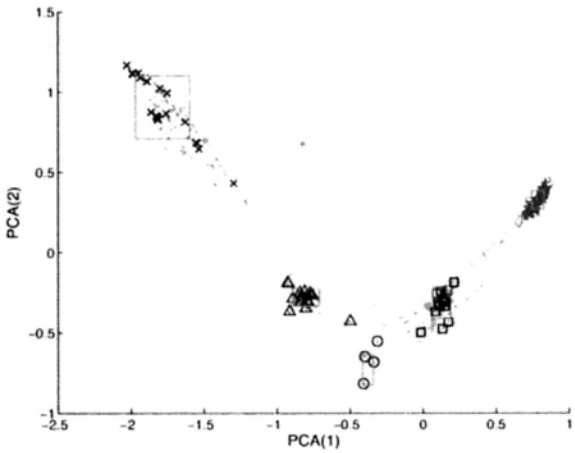
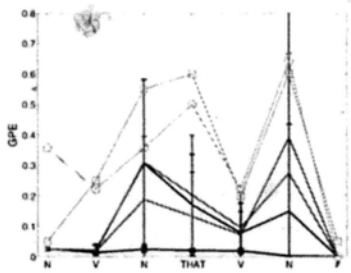
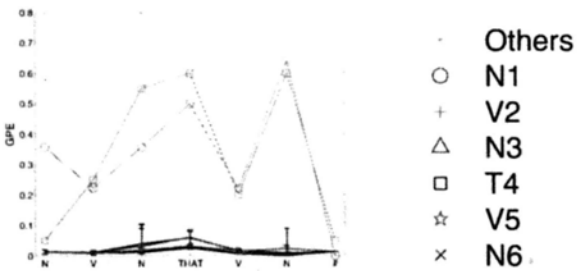
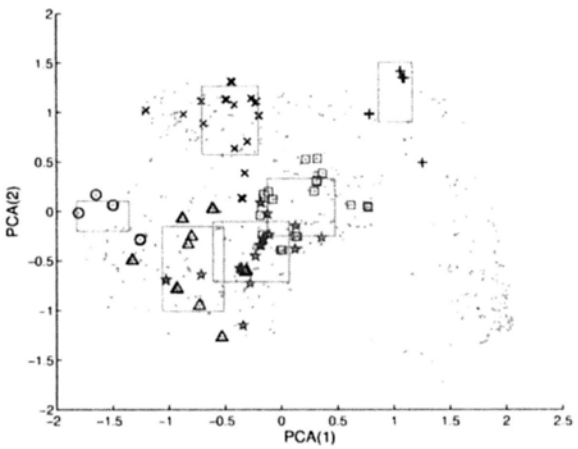
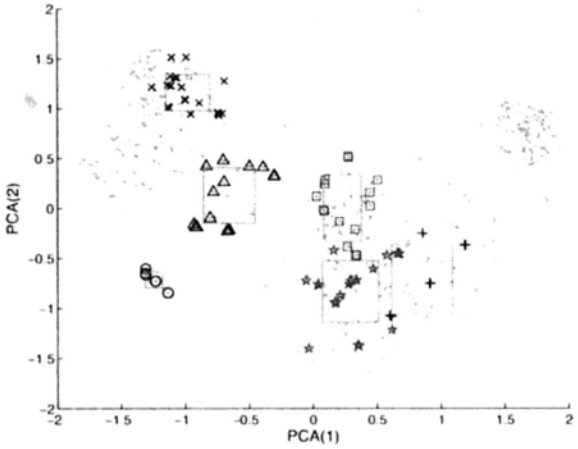
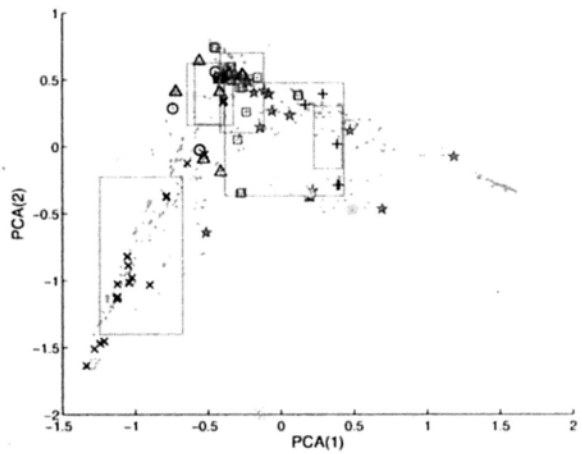
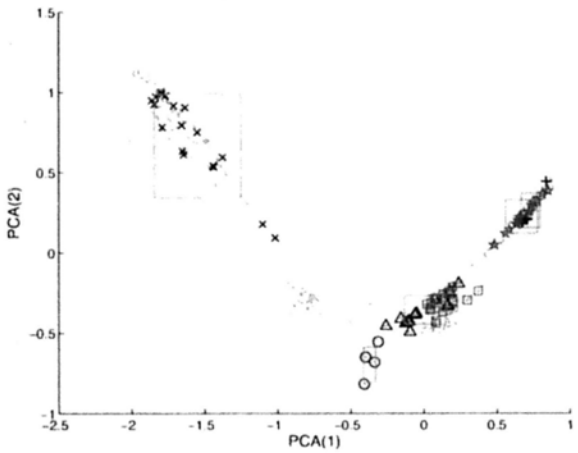
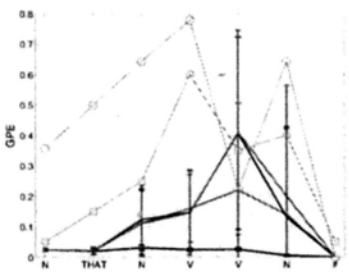
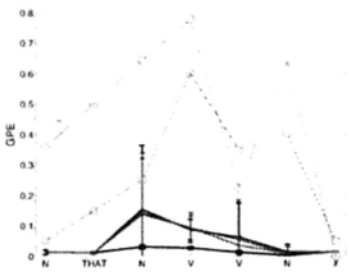
| Network #5, **high** GPE attained | Network #8, **low** GPE attained |
|---|---|
| (a) 1st hidden layer | (b) 1st hidden layer |
|  |  |
| (c) 2nd hidden layer | (d) 2nd hidden layer |
|  |  |
| GPE evaluation | GPE evaluation |
|  |  |

Table 4.9: Hidden layer activations during the processing of $M = 4$ **centre-embedding sentences** obtained from a 2-hidden-layer-SRN that was least successful in generalisation (network #5) and a 2-hidden-layer-SRN that was most successful in generalisation (network #8), see text for details.

| Network #5, **high** GPE attained | Network #8, **low** GPE attained |
|---|---|
| (a) 1st hidden layer | (b) 1st hidden layer |
|  |  |
| (c) 2nd hidden layer | (d) 2nd hidden layer |
|  |  |
| GPE evaluation | GPE evaluation |
|  |  |

A qualitative difference between the distribution of hidden layer activations obtained from the first hidden layer and that obtained from the second hidden layer seems to have emerged for both SRN #5 and SRN #8. Clustering according to word class, i.e. nouns versus verbs, was observed in both hidden layers, but in the second hidden layer, the clustering was more dependent on sentence position. This trend was more obvious for successful networks, i.e. SRN #8, (b) and (d) in Table 4.8 (p. 76), that showed greater ability to generalise.

For network #8, in processing testing set right-branching sentences, Table 4.8 (p. 76), non-overlapping categories were formed in the *second hidden layer* distinguishing the main clause subject (N1), main clause object / relative clause subject (N3) and relative clause object (N6). This agrees with the low GPE evaluation obtained by this 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 the verbs, V1 and V2, almost completely overlapped with one another. This was due to the limitation of the task in the paradigm, as in both of these two sentence positions only nouns could be the correct continuation. The task did not require the network to make further distinction.

In processing testing set centre-embedding sentences, Table 4.9 (p. 77), a similar pattern was also present. N3 and V4 words overlapped with one another as in both sentence positions verbs were the grammatical continuation. This was true also for T2 and V5 as both should predict nouns to follow. Notice that sentence position was not the sole information governing the clustering, as N3 in right-branching sentences

and N3 in centre-embedding sentences formed two distinct groups, cf. blue triangles in panel (d) of Table 4.8 and Table 4.9.

The pattern of clustering of the hidden layer activation in the case of less successful networks, i.e. SRN #5, (a) and (c) in Table 4.8 (p. 76), differed from that of successful networks in a sense that clear boundaries between groups were absent. In the first hidden layer, though nouns and verbs did form their own groups but the clustering was unlike that formed in successful networks. The clustering in successful networks was more *categorical* as demonstrated by a larger between groups distance in PCA space compared to the within group distance in PCA space. In the second hidden layer, the difference was even more transparent. Less successful networks were not able to develop the hidden layer activations distinguishing nouns in different contexts.

For completeness, Table B.1 (p. 130) and Table B.2 (p. 131) in Appendix B plots the hidden layer activations from two more networks, one was the second least successful and the other was the second most successful in generalisation.

## 4.8 Summary of Chapter 4

The dismissal of SRNs as a model for cognition by van der Velde was based on a premature analysis of the networks' performance and is therefore unconvincing. The experimentations with SRNs, under the framework of assessing combinatorial productivity exhibited by SRNs, 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

(iii) networks with two hidden layers show better ability to generalise

The analysis on SRNs' hidden layer activations showed that different layers could play differential roles in connectionist networks for language modelling. Among the 2-hidden-layer-SRNs I 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 sentence position in the second hidden layer. My speculation that the success of 2-hidden-layer-SRNs was driven by the categorization on top of general noun-verb distinction was supported by such an observation, particularly from the contrast between the more successful networks with the less successful ones.[19]

[19]In the simulations reported in this thesis, the subject relativisation of right-branching (N-V-N-that-V-N) and the object relativisation of centre-embedding (N-that-N-V-V-N) were used as the training and testing materials. One might ask to what extent the argument will apply if the object relativisation of right-branching (N-V-N-that-N-V) and the subject relativisation of centre-embedding (N-that-V-N-V-N) were used instead.

My analysis showed that success of generalisation depends on how well the network captures the sentence structure which is manifested as the serial order of words. The use of an alternative artificial language (i.e. with N-V-N-that-N-V and N-that-V-N-V-N) would actually make the training and generalisation task *easier* for the networks. Since in this language, the information about absolute sentence position plus the word class information about the current word already uniquely identify sentence type and hence uniquely determine what words can grammatically follow. Unlike the current setting of the language that requires an additional information about the left context for successful prediction. In short, the use of the alternative language would make the generalisation task less taxing and therefore it is expected that SRNs will not degrade in their performance on generalisation.

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, I see no reason why that cannot happen in the living brain.

# Part II

# Language Processing in the Real Brain

# Chapter 5

# Event-related Brain Potential Study on Reading

## 5.1 Overview—Finding early brain signature to semantic processing

In this chapter, I will report an event-related electrical potential (ERP) study[1] conducted to investigate early neurophysiological response to contextual influences in reading. Participants were instructed to read and comprehend visually presented Chinese sentences while their ERPs to critical words were recorded. Both the classic N400 component and an early ERP component, the P200, were found to be sensitive to the experimental manipulations on semantic congruency. A reduction in amplitude

[1] A portion of the findings discussed in this chapter has been submitted to *Neuropsychologia*, (Wong & Wang)

of the P200 to incongruent critical words was observed and it was lateralised to the left hemisphere. My results provide evidence that within the first 200 ms the brain has already started to process the lexical meanings of a word and this process interacts with the information carried by the sentence context. Such an early brain signature to semantic processing poses a challenge to the serial, syntax-first model of language processing. Implications for theories of Chinese reading will also be discussed.

## 5.2 Background

The brain mechanisms involved in extracting information from incoming linguistic signals, be they in spoken or in written forms, have long been a central theme of cognitive research (Marslen-Wilson & Tyler, 1975). It has been acknowledged that to make sense of the continuous stream of incoming signals, different brain processes are involved to rapidly reveal information that is made available at different levels (Bentin, Mouchetant-Rostaing, Giard, Echallier & Pernier, 1999; Lau, Phillips & Poeppel, 2008; van den Brink, Brown & Hagoort, 2001). As for reading, most theories admit the existence of functionally distinct brain processes. These processes start from the perceptual (or sub-lexical) level where the brain is engaged to convert visual inputs to orthographic, phonological and/or semantic information. This is followed by processes operating at a lexical level that extract meanings of the incoming words. Based on the lexical information then available, sentence level operations that attempt to integrate the lexical meaning into the running sentence context take place to construct a coherent "semantics".

However, the physiological manifestations of these processes and the mapping of the time course of the early *lexical access* and of the later *semantic integration* remains to be a subject of current research. The event-related potential (ERP) study of van den Brink et al. (2001), for instance, reported that the amplitude of the N200 ERP component, negativities with a latency interval of 140–180 ms, to auditorily presented words that were incongruent with the sentence contexts were larger than that to congruent words. The authors argued that the N200 reflects an initial assessment *on the semantic features* of those lexical candidates activated by perceptual level processing. They proposed that this initial assessment is based on how well the lexical candidates semantically fit into the sentence context, and from this it was concluded that contextual influences during auditory sentence comprehension is in action even during such a early time frame, consistent with a cascade model of sentence comprehension (Bentin et al., 1999; Hauk, Ford, Pulvermüller & Marslen-Wilson, 2006; Ruz & Nobre, 2008).

On the other hand, a more classic view of sentence comprehension predicts that semantic integration does not take place until a much later time frame, as indexed by the N400 component. The N400 effect was first reported in the study of Kutas & Hillyard (1980) in which ERPs to words that introduced semantic incongruency with the sentence contexts were extracted, for example, the word "*socks*" in the sentence "*he spread the warm bread with socks*". The N400 ERP component, negativities in the time region around 300–500 ms after stimulus onsets elicited by these words was found to be larger, more negative, in amplitude than those to congruent words, e.g.

"*butter*". Subsequent studies collectively concluded that the amplitude of N400 is inversely proportional to a word's cloze probability, the empirical likelihood that the word completes a partial sentence. The amplitude of the N400 is generally accepted as an index of the ease of contextual integration of the lexical meaning of a word with the sentence context (Hagoort, Hald, Bastiaansen & Petersson, 2004; Kutas & Federmeier, 2000, but see Lau et al., 2008).

Under this *late semantic integration framework*, the observed early ERP effects to semantic incongruency, the N200 and N270 effects in the study of van den Brink et al. (2001) and Connolly & Philips (1994), respectively, are better explained as a mismatch between the contextually expected perceptual features with the actual perceived features. In the case of auditory sentence comprehension, Connolly & Philips (1994) referred to these responses as the phonological mismatch negativities. This late integration view is exemplified by Friederici's (2002) model of language processing, in which auditory sentence comprehension is divided into temporally ordered phases of processing:

(i) Identification of phonemes, as indexed by N100 (at around 100 ms after stimulus onset);

(ii) Identification of syntactic categories, as indexed by early left-anterior negativity LAN (at a time window of 150–200 ms);

(iii) Morphosyntactic and lexical-semantic processing, as indexed by LAN and N400, respectively (at a time window of 300–500 ms);

(iv) Integration of the information from different levels of analysis, as indexed by P600 (at around 600 ms).

None of the ERP components in a time range earlier that the N400 is said to be related to lexical semantic processing. The implication of the model is that contextual integration does not take place earlier than the time range indexed by N400.

What is even more controversial is the important but equivocal premise behind the dominant late integration view on sentence comprehension, namely, the lack of an early ERP component that indexes lexical semantic processing. Compared to the overwhelming discussion on N400 since Kutas & Hillyard (1984, 1980), ERP studies that reported the possibilities of early components of this kind emerged only recently. Apart from the auditory N200 (van den Brink et al., 2001), studies that utilised the classical visual word-by-word reading for comprehension paradigm reported that a P200 component may index access to lexical semantics (Penolazzi, Hauk & Pulvermüller, 2007; Meng, Tian, Jian & Zhou, 2007). A similar P200 modulation in a semantic relatedness judgement task on word pairs was also reported (Landi & Perfetti, 2007). Studies on word recognition potential also suggested a potential early brain response to semantic processing (Hinojosa, Martín-Loeches, Muñoz, Casado & Pozo, 2004; Martín-Loeches, Hinojosa, Casado, Muñoz & Fernández-Frías, 2004). However, the early brain response described in these papers was a negativity that peaked at around 250 ms after stimulus onset. Thus, it exhibited the opposite polarity to the P200 and a longer latency.

The sparseness of electrophysiological data on early in-context word processing might partly be accounted for by the observation that early components are generally of smaller amplitudes and thus likely to be overlooked, as suggested by Hauk et al. (2006). The ERP study of Penolazzi et al. (2007) echoed such a view. Penolazzi et al. (2007) reported that the amplitudes of ERPs to visually presented critical words in the time windows 110–130 ms and 170–190 ms indeed varied with the words' cloze probabilities in a sentence comprehension task. However, the effects of semantic congruency were not additive in these early time windows; instead, it interacted with word length. The interaction effect between other linguistic factors, frequency of usage and word length, was also found to be significant in the two time windows. Based on these findings, they proposed that other studies failed to detect an early ERP signature to semantic processing because these relevant linguistic factors were not systematically controlled.

In light of the speculations of Hauk et al. (2006) and Penolazzi et al. (2007), I conducted an ERP study to investigate the presence of an early brain signature to semantic processing during reading. I attempted to elicit such a response by contrasting ERPs to words that are either congruent or incongruent with the preceding sentential contexts. I specifically designed the materials aiming to achieve a higher sensitivity. The key innovation in my design was that I use the exact same set of critical words in both the congruent and the incongruent conditions, such that factors like the frequency of usage, the visual complexity of the critical words (e.g. number of strokes) and their lexical meanings were all balanced. By doing so, I kept the contamination of

experimental effects due to stimulus variability to a minimum with the aim to detect early ERP effects that might escape notice otherwise.

## 5.3 Methods

### 5.3.1 Participants

Thirty-three students from The Chinese University of Hong Kong, who are native speakers of Hong Kong Cantonese, participated in the experiment. Participants were compensated at a rate of about HK$250 and their written informed consent was obtained. The study was approved by the Survey and Behavioural Research Ethics Committee of The Chinese University of Hong Kong. All participants were right-handed[2] (Snyder & Harris, 1993; Tan, Spinks, Gao, Liu, Perfetti, Xiong, Stofer, Pu, Liu & Fox, 2000) with normal or corrected-to-normal vision, and reported no history of neurological illness. Electroencephalograph (EEG) recordings from thirteen participants were excluded from further analysis due to excessive head movement and poor fitting of the sensor nets (three of them) and contamination of ocular artefacts. Data from twenty participants (ten males; mean age: 20.85 years, s.d.: 2.3 years) was included in the final analysis.

[2]The handedness questionnaire used was attached in Table C.2 (p. 140) in Appendix C

### 5.3.2 Materials

#### Congruent sentences

The composition of congruent sentences was based on 96 sentences extracted from the Hong Kong Parallel Text corpus (Ma, 2004) in the Linguistic Data Consortium. Two of the three sub-corpora were used, namely, Hong Kong Hansards (HKH) Parallel Text and Hong Kong News (HKN) Parallel Text. HKH contains "*excerpts from the Official Record of Proceedings (hansards) of the Legislative Council of the HKSAR*" (Ma, 2004) and HKN contains press releases from the government of HKSAR. For both of the sub-corpora, some text were translated from Chinese to English and some text were translated from English to Chinese.

The extracted sentences were in the form semi-formal written Cantonese, a form which commonly appears in newspapers and textbooks. I simplified the extracted sentences and used them as the congruent sentences. Simplification was done mainly to reduce the sentence length and to remove punctuations while not hindering the naturalness and readability of the sentences. Table 5.1 (p. 92) lists five sample sentences to illustrate the simplification process. For each sentence in Table 5.1:

- The first two rows were extracted from the corpus as they were
- The third row was the simplified and segmented version that would eventually be delivered as experimental stimuli to the participants

- The fourth row was the English gloss
- The last row was the translation in English

Sentences 1, 4 and 5 involved simplification from long sentences to short ones with the removal of punctuations. Sentences 2 and 3 involved re-phrasing to improve readability. As the stimuli would be presented in a word-by-word manner to the participants, the sentences were segmented manually into words. Most words were di-syllabic, hence appeared as a two-character word, some of them were mono-syllabic and tri-syllabic. Only three of them were words of four-syllable long.

**Incongruent sentences**

The incongruent sentences were constructed by exchanging the critical words[3] between pairs of congruent sentences, subject to the constraint that the critical words were of the same syntactic category as judged by their usage in the sentences. For example, in Table 5.2 (p. 93), $S_1'$ was the incongruent version of $S_1$ and it was composed by replacing the noun critical word "*hill-fire*" of $S_1$ with the critical word "*drug*" from $S_2$, which was also used as a noun in $S_2$. Likewise $S_2'$ contained the left and right context of $S_2$ but with the critical word replaced by "*hill-fire*" from $S_1$. The total of 192 experimental sentences were divided into two complementary sets. One set contained the congruent versions of a member of those sentence pairs and the incongruent versions of the other

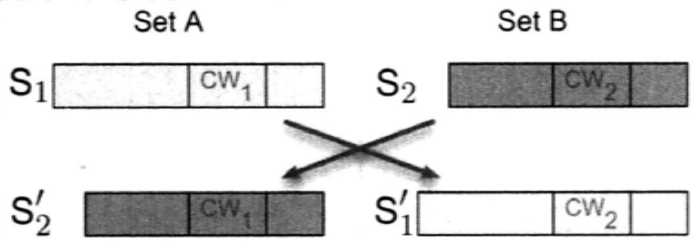
[3] I sampled randomly sentences from the corpus in a pair-by-pair manner. I identified if critical words can be found such that semantic anomalies would be introduced once their positions were swapped. If no appropriate critical word pairs were identified, I discarded the sentence pairs.

Table 5.1: An illustration of simplification and segmentation of sentences extracted from the corpus

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | From corpus (Chinese) | 目前的十多萬失業人士中，半數只有初中或以下教育程度。 | | | | | | | | |
| | From corpus (English) | Of the 100,000-odd unemployed at the moment, half are educated to junior secondary level or below. | | | | | | | | |
| | Chinese | 失業 | 人士 | 中 | 半數 | 只有 | 初中 | 以下的 | 教育 | 程度 |
| | Gloss | unemployed | people | among | half | only-have | junior-secondary | below | education | level |
| | English | Of the unemployed, half are educated to junior secondary level or below | | | | | | | | |
| 2 | From corpus (Chinese) | 重陽節前後是傳統發生山火的高危季節 | | | | | | | | |
| | From corpus (English) | The time of the year around Chung-Yeung Festival is traditionally a high-risk period for hill fires. | | | | | | | | |
| | Chinese | 重陽節 | 前後 | 是 | 發生 | 山火 | 的 | 高危 | 季節 | |
| | Gloss | CYF | around | is | to-happen | hill-fire | AS | high-risk | season | |
| | English | The time of the year around Chung-Yeung Festival is a high-risk period for hill fires | | | | | | | | |
| 3 | From corpus (Chinese) | 青少年藥物濫用者有所增加 | | | | | | | | |
| | From corpus (English) | An increase was observed in the number of young drug abusers | | | | | | | | |
| | Chinese | 青少年 | 濫用 | 藥物 | 的 | 情況 | 有所 | 增加 | | |
| | Gloss | Youngster | to-abuse | drug | AS | circumstances | there exists | increase | | |
| | English | An increase was observed in the number of young drug abusers | | | | | | | | |
| 4 | From corpus (Chinese) | 假如他們出現非典型肺炎病徵，會被送往醫院接受治療及隔離。 | | | | | | | | |
| | From corpus (English) | If they have developed symptoms of the disease, they will be admitted to hospital for isolation and treatment. | | | | | | | | |
| | Chinese | 他們 | 出現 | 肺炎 | 病徵 | 被 | 送往 | 醫院 | 接受 | 治療 |
| | Gloss | they | develope | pneumonia | symptom | is | admitted | hospital | receive | treatment |
| | English | they have developed symptoms of pneumonia and were admitted to hospital for treatment | | | | | | | | |
| 5 | From corpus (Chinese) | 所以，我們認為居屋本身是有助紓緩政府對公營房屋的財政負擔。 | | | | | | | | |
| | From corpus (English) | As such, we hold that the HOS is helpful to alleviating the public housing-related financial burden on the Government. | | | | | | | | |
| | Chinese | 居屋 | 有助 | 紓緩 | 政府的 | 財政 | 負擔 | | | |
| | Gloss | HOS | helpful | alleviate | of-government | financial | burden | | | |
| | English | Home Ownership Scheme is helpful to alleviating the financial burden on the Government | | | | | | | | |

Table 5.2: Samples of experimental materials used. $S_1$ and $S_2$ are congruent sentences. Figure at the bottom: by exchanging the critical words of this sentence pair, incongruent versions ($S'_1$ and $S'_2$) were formed. The experimental sentences were divided into two complementary sets, denoted as Set A and Set B. Participants were presented with either one of them.

| | Word 1 | Word 2 | Word 3 | Word 4 | Word 5 | Word 6 | Word 7 | Word 8 |
|---|---|---|---|---|---|---|---|---|
| $S_1$: | 重陽節 | 前後 | 是 | 發生 | 山火 | 的 | 高危 | 季節 |
| Gloss | Chung-Yeung Festival | around | is | to-happen | hill-fire | AS | high-risk | season |
| English | The time of the year around Chung-Yeung Festival is a high-risk period for hill fires. | | | | | | | |
| $S_2$: | 青少年 | 濫用 | 藥物 | 的 | 情況 | 有所 | 增加 | |
| Gloss | Youngster | to-abuse | drug | AS | circumstances | there exists | increase | |
| English | An increase was observed in the number of young drug abusers. | | | | | | | |
| $S'_1$: | 重陽節 | 前後 | 是 | 發生 | 藥物 | 的 | 高危 | 季節 |
| Gloss | Chung-Yeung Festival | around | is | to-happen | drug | AS | high-risk | season |
| English | The time of the year around Chung-Yeung Festival is a high-risk period for drugs. | | | | | | | |
| $S'_2$: | 青少年 | 濫用 | 山火 | 的 | 情況 | 有所 | 增加 | |
| Gloss | Youngster | to-abuse | hill-fire | AS | circumstances | there exists | increase | |
| English | An increase was observed in the number of young hill fires abusers. | | | | | | | |




AS: adjectival suffix
Blue: Congruent critical word
Red: Incongruent critical word

member of those sentence pairs (formally, Set A = $\{S_x, S'_y\}$; Set B = $\{S'_x, S_y\}$, for $x = 1, 3, 5, ...95$; $y = 2, 4, 6, ...96$). Participants were randomly assigned to either set of the experimental sentences.

The full set of materials used for the experimental conditions are given in Table C.1 (p. 133) in Appendix C.

I employed such a paradigm to construct my materials such that the exact same set of critical words were used in both the congruent and the incongruent condition. Factors other than the mismatch with the sentence context, such as word frequency, visual complexity of the words, and even their lexical meaning were automatically balanced across the two experimental conditions.

Among the 96 congruent-incongruent sentence pairs, 58 of them contained a noun as the critical word, 24 of them contained a verb as the critical word, and 14 of them contained an adjective as the critical word. Overall, among the set of words in the 96 congruent sentences, 72.8% (567 out of 790) were di-syllabic words; 14.3% were mono-syllabic; 13.5% were tri-syllabic; and only three of them were four-syllable words. All critical words were di-syllabic words and they appeared on the screen as two Chinese characters (see Table 5.2). Only ERPs to noun critical words will be discussed in this study as it has been established that nouns and verbs elicit ERPs with distinct patterns (Hauk & Pulvermüller, 2004; Hauk, Johnsrude & Pulvermüller, 2004). The inclusion of verb and adjective critical words was aimed to make the occurrences of anomalies less predictable. Along the same line, 48 of 63 fillers sentences, which

were also congruent sentences extracted from the same corpus, were mixed with experimental materials, the other 15 were used as practise trials. The average sentence length for the 58 experimental sentence pairs was 7.9 (s.d.: 1.9). On average, a critical word appeared as the fifth or the sixth word (mean: 5.6; s.d.: 1.8) in a sentence. I did not fix the sequential position of the critical words nor the sentence structures because I weighed naturalness and unpredictability of the occurrence of the anomalous words over the ERP variations due to sentence position and sentence structure. Nevertheless, such variability was balanced across the two conditions.

### 5.3.3 Task

Participants were seated in a quiet room in font of a LCD screen placed at a distance of about one meter away from them. They were instructed to read silently sentences that would be presented on the screen and to make a binary judgement on the overall acceptability of each of the sentences when prompted. To make the reading task as naturalistic as possible and to minimise cognitive load due to meta-linguistic analysis, participants were told that their responses need only to reflect their own impression of the materials. Instruction sheet delivered to participants before the start of each EEG session was attached in Table C.3 (p. 141) in Appendix C. Verbal description in Cantonese was given after the participants had read the instruction sheet.

Sentences were presented on the screen in a word-by-word manner and were displayed in white colour against a black background in Arial Unicode MS font. Each Chinese character was about 1.5 cm in width and in height. The experimental control

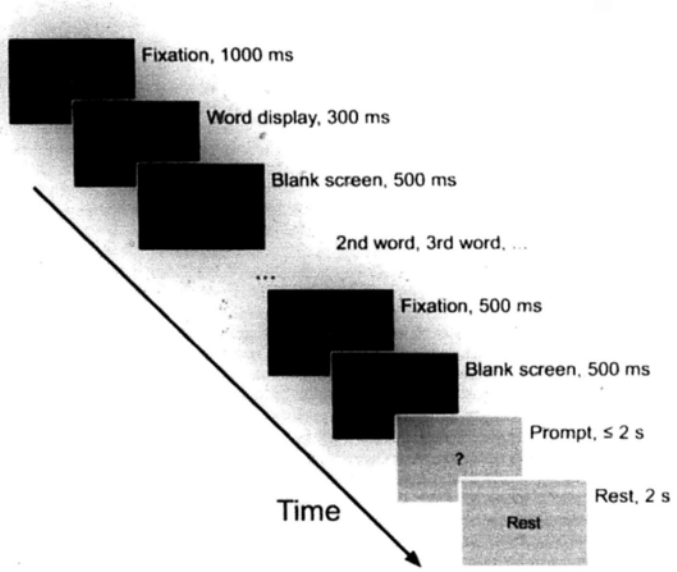



Figure 5.1: Word-by-word presentation of a sentence in a trial. Participants were instructed to maintain their eye gaze at the centre of the screen and reduce eye blinks during sentence presentation, as indicated by a black background.

was done by E-Prime (version 1.2 by Psychology Software Tools, Inc.) running in a desktop computer and the stimuli were displayed with the screen resolution and the refresh rate set to $1280 \times 1024$ pixel and 60 Hz, respectively.

A computer mouse was used as the respond device and participants held it with both hands. Behavioural responses were made by pressing either the left mouse button with the left thumb or the right mouse button with the right thumb. Participants were instructed to maintain their eye gaze at the centre of the screen, where stimuli would appear, and try to blink less frequently during sentence presentation.

Each trial, as depicted in Figure 5.1 (p. 96), was started with a fixation screen for one second after which the first word of the sentence appeared for 300 ms followed by a blank screen that lasted for 500 ms. The remaining words of the sentence were displayed in the same manner, i.e. a word slide followed by a blank screen as a visual

mask. The fixation screen was shown again for 500 ms, followed by a blank screen that lasted for 500 ms, to indicate the end of the sentence. Participants were then prompted, with a question mark "?" displayed on the screen for at most two seconds, for their ratings on the acceptability of the sentence. They were given a two-second break between every trial and they were advised to rest and blink their eyes during these breaks.

The set of 144 sentences (48 congruent, 48 incongruent and 48 fillers) were randomly divided into ten blocks, nine of which contained 15 sentences. For the first five blocks, a "Yes, acceptable" response was designated by pressing the left mouse button and a "No, not acceptable" response was designated by pressing the right mouse button. The button assignment was reversed for the remaining five blocks. The order of sentence presentation within each block was randomised for each participant. The order of the blocks was also randomised for each participant but they were always exposed to the blocks that required a "Yes, acceptable" response with the left mouse button press first. Each block lasted for about three minutes and participants were instructed to close their eyes and rest during the breaks administered between each block. A practice block, with 15 filler sentences that did not appear in the actual experimental blocks, was delivered before the first session. As participants were told to withhold their responses until when prompted, the reaction time data will not be analysed. Instead, a comparison of participants' acceptability ratings and my classification of sentence congruency was made to confirm that the participants were indeed performing the task of reading for comprehension.

### 5.3.4 EEG recording

Electroencephalograph (EEG) data were recorded using Geodesic EEG System 250 (Electrical Geodesics, Inc.) with 128-channel Ag/AgCl electrode arrays. The approximate locations of each electrode was drawn in Figure 5.2 (p. 99)[4] in which the labelling of the electrodes follows the international 10/10 system (Chatrian, Lettich & Nelson, 1985) except for electrodes labelled with an "EGI-" prefix.

Recordings were done on a separate computer running NetStation (version 4.2, Electrical Geodesics, Inc.). Data were recorded at a rate of 1000 Hz, referenced to the vertex, filtered with an analogue band-pass filter (0.1 Hz to 400 Hz), and digitised using a 16-bit A/D converter. The EEG were re-referenced offline against average-mastoid reference, i.e. the mean voltage of the left and right mastoid electrodes was used as the reference potential, and filtered with a digital low-pass filter with cut-off frequency set at 40 Hz. Eye blinks were monitored with vertical electrooculograph (EOG) electrodes placed above and below each eye whereas eye movements were monitored by horizontal EOG electrodes placed near the outer canthi. The electrode impedances were generally kept below 60 kΩ (amplifier input impedance was 200 MΩ).

### 5.3.5 Artefact detection

EEG segments to critical words of the experimental sentences were extracted from 200 ms before the stimulus onset to 800 ms after it. The mean voltage in the 50 ms

[4]The figure was plotted using the software package BESA, Brain Electrical Source Analysis (http://www.besa.de)

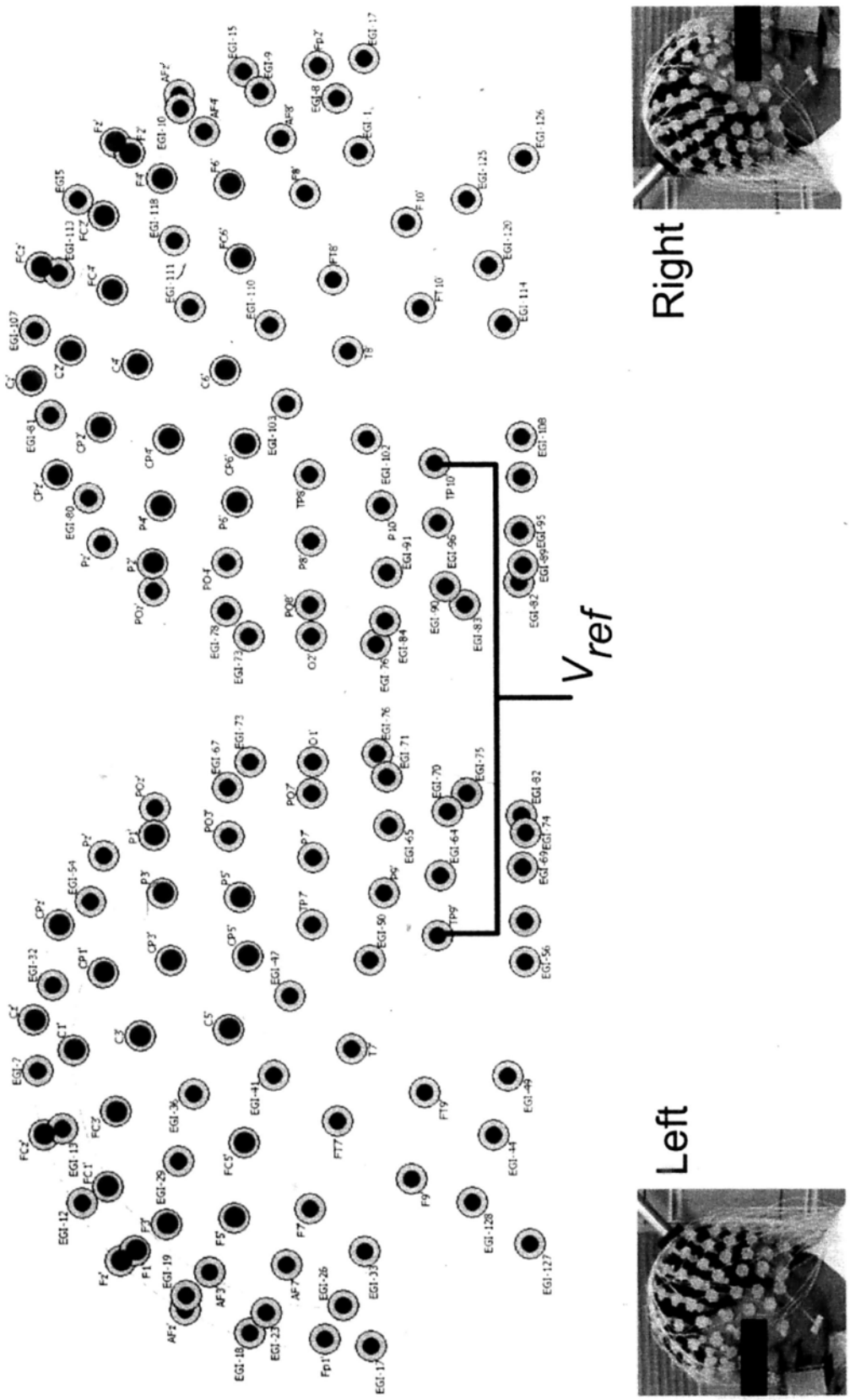



Figure 5.2: Electrode layout of the 128-channel EEG system used in this study. The labelling of the electrodes follows the international 10/10 system except for electrodes with an "EGI-" prefix as they are outside the 10/10 system, which covers 81 channels only.

interval prior to stimulus onset was used as the baseline in subsequent ERP derivation. The EEG segments were subjected to artefact detection. Electrodes with readings[5] varied greater than 200 μV were marked as bad channels, data from these channels were replaced by spherical spline interpolation. Ocular artefacts were detected by analysing the EOG segments with a moving window of 640 ms in width running along the segments. If the readings[6] from a trial varied greater than 140 μV in a vertical EOG, the trial was marked as contaminated by eye-blink artefacts. Likewise if the readings varied greater than 55 μV in a horizontal EOG, the trial was marked as containing eye-movement artefacts. Trials with ocular artefacts were excluded from averaging to ERP. Moreover, if more than 15% of the experimental trials from a participant were to be excluded, all EEGs from such participants, as mentioned, ten of them, were excluded from further analysis. For the remaining twenty-subject data, on average 94.4% (minimum: 86.2%) of the experimental trails were free of ocular artefacts.

## 5.4 Results

### 5.4.1 Behavioural response

In general, participants' acceptability rating agreed with my classification of the experimental sentences. On average, 83.4% (s.d.: 16.2%) of the congruent sentences were rated as "acceptable" and 86.9% (s.d.: 14.5%) of the incongruent sentences were rated as ,"not acceptable". As a baseline, 85% (s.d.: 15.2%) of the filler sentences were

[5]Smoothed by moving average with a moving time window of 80 ms in width
[6]Also smoothed by moving average with a 80 ms wide time window

rated as, "acceptable". The behavioural data confirmed the validity of the task as well as the design of the materials.

### 5.4.2 Electrophysiological response

Figure 5.3 (p. 102) shows the grand average ERPs, time-locked to the onset of the critical words, from a selected subset of the international 10/10 system, representatives from the four quadrants (left-anterior: F3; right-anterior: F4; left-posterior: P3; right-posterior: P4) plus midline and central electrodes (Fz, Cz, Pz, C3 and C4).

Figure 5.4 (p. 103), on the other hand, plots the ERPs obtained from 25 more electrodes to provide a more completed representation of the data I obtained. Furthermore, ERPs from 24 electrodes in Figure 5.4 will be included in the subsequent statistical analysis to be discussed in the next section. The locations of these 34 electrodes were marked with blue dots in Figure 5.2 (p. 99).

Visual inspection of the waveforms reveals two time windows of interest in which experimental effects on ERPs were observed. ERPs to congruent critical words and incongruent critical words started to diverge at around 150 ms after stimulus onset, during which the ERP component was a positive deflection. The difference in waveforms between the two conditions remained salient up to 200 ms. I refer to this time region as the P200 region. This definition of the P200 component is consistent with those in ERP studies on English sentence comprehension (Penolazzi et al., 2007) and word reading (Landi & Perfetti, 2007) as well as Chinese word reading

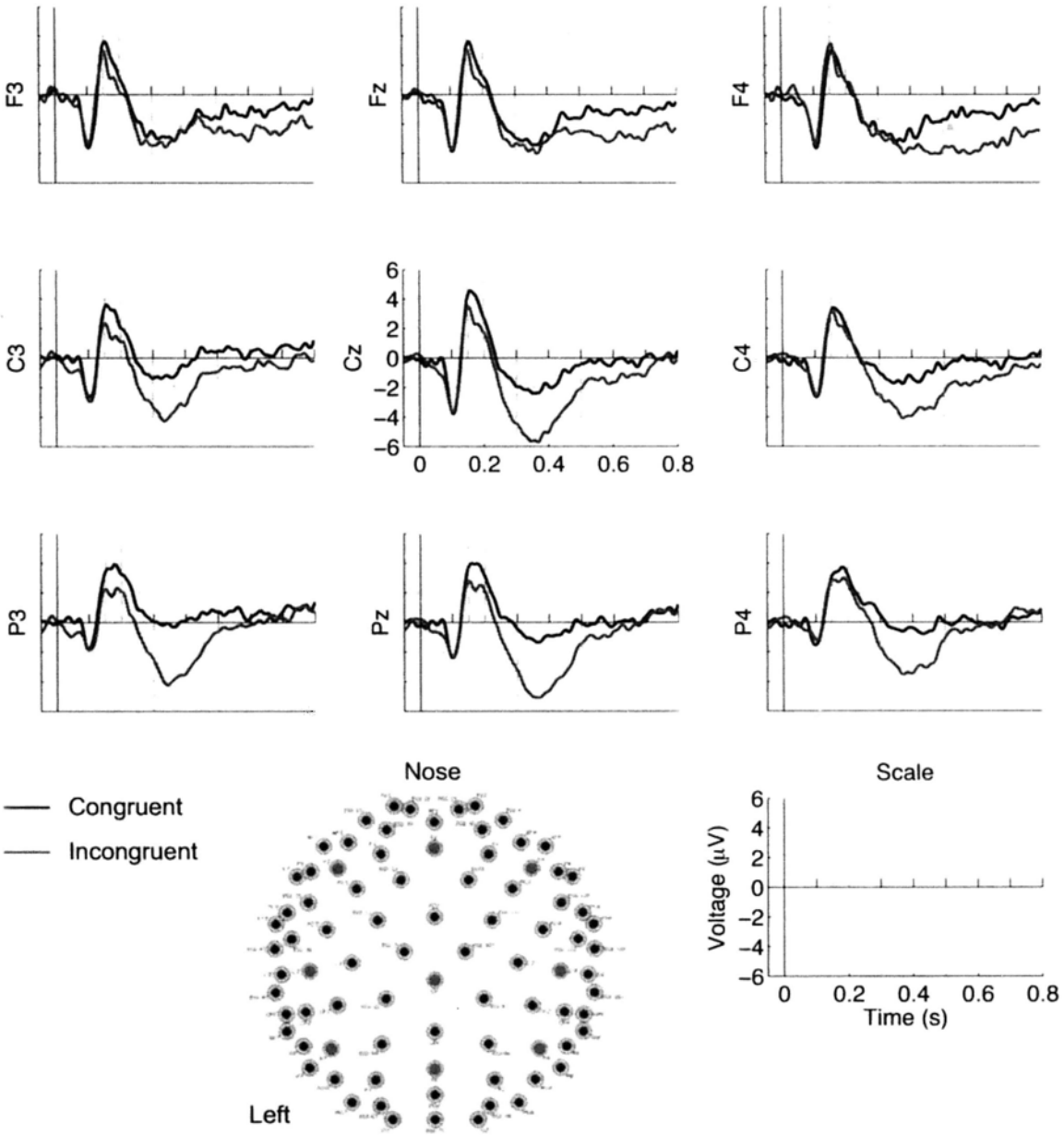



Figure 5.3: Grand average ERPs from nine selected 10/10 electrodes. Red lines denote ERPs to critical words that were incongruent with sentence contexts. Blue lines denote ERPs to congruent critical words. Time zero of the ERPs corresponds to the onset of the critical words. Vertical dashed lines denote the onsets and offsets of the time windows of interest, P200 region: [150 ms, 200 ms]; N400 region: [300 ms, 500 ms]. Re-reference montage: average-mastoid; baseline interval: [−50 ms, 0 ms). The figure at the bottom gives a top view of the electrode placement in which a red dot denotes an electrode from which the ERPs were obtained.

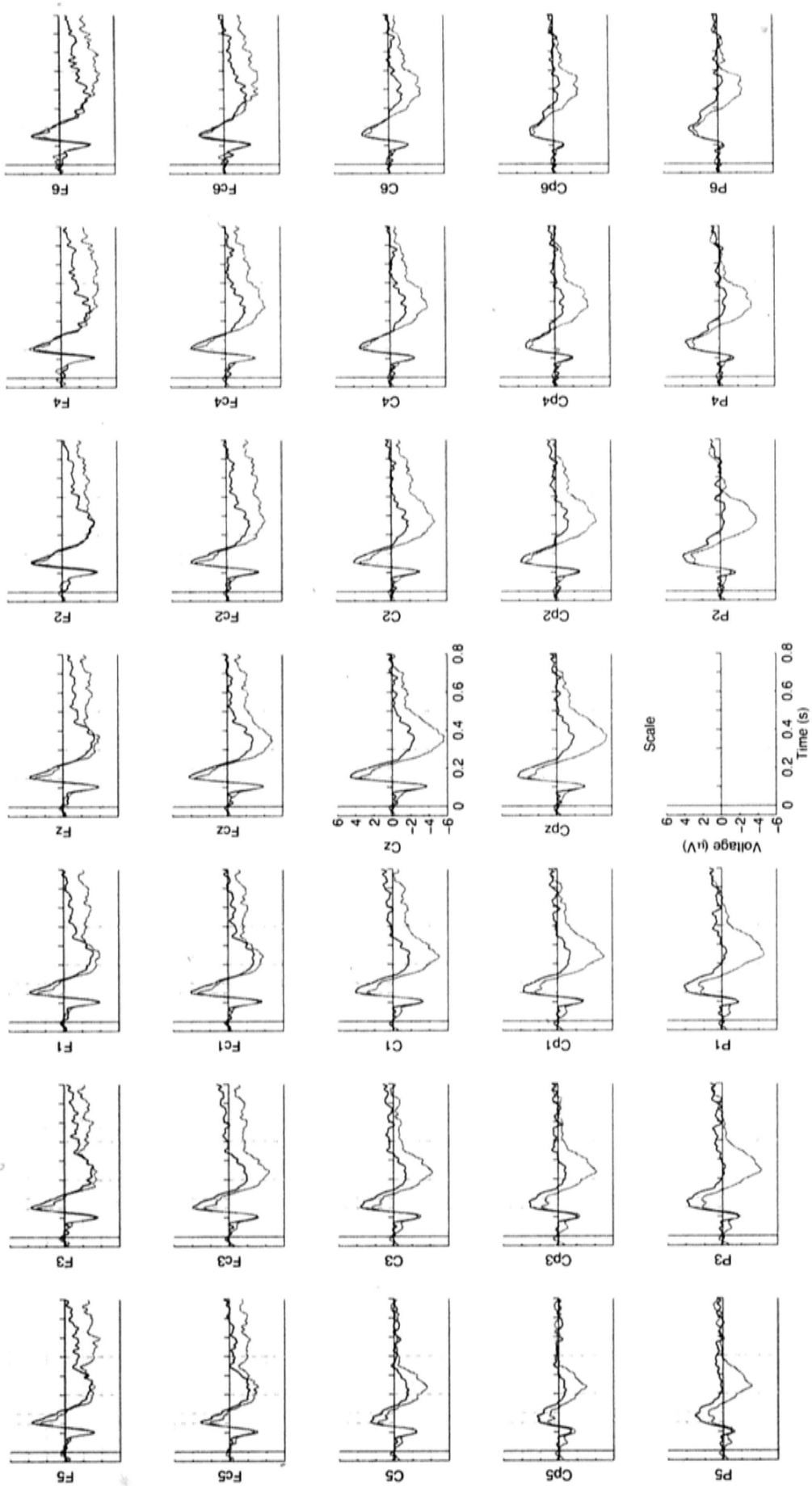



Figure 5.4: Grand average ERPs from 34 selected 10/10 electrodes

(Liu, Perfetti & Hart, 2003). The classic N400 response, negativities in the 300 ms to 500 ms region (Kutas & Hillyard, 1980; Kutas & Federmeier, 2000), to words was also observed in the grand average waveforms, and it was also sensitive to the experimental manipulations.

## 5.5 Analysis

The mean amplitudes of the ERPs in these two time windows (150–200 ms for P200; 300–500 ms for N400) were extracted as the dependent variable for statistical analysis. Analysis of variance (ANOVA) was first conducted on representative electrodes (F3, F4, P3 and P4) of the four quadrants in which a repeated measure $2 \times 2 \times 2$ ANOVA was conducted for each time window with CONGRUENCY (congruent-incongruent), HEMISPHERE (left-right), and LOBE (anterior-posterior) as the three within-subject factors. To make use of the high-density recording and to validate my findings with data that covered a wider region of the scalp, the ANOVA was conducted also on ERPs after regional-averaging. For each region (HEMISPHERE $\times$ LOBE), ERPs from 6 electrodes were averaged and fed to ANOVA. Table 5.3 (p. 105) lists the 24 electrodes included in this *regional-averaging ANOVA*.

### 5.5.1 The P200 (150–200 ms)

Inspection of the grand average ERP waveforms, Figure 5.3 (p. 102) and Figure 5.4 (p. 103), reveals that the brain activities showed an early sensitivity to semantic

Table 5.3: Electrodes from which ERPs were averaged by region and submitted to regional-averaging ANOVA. Four regions were defined according to dimensions LOBE (anterior-posterior) and HEMISPHERE (left-right). Bolded electrode names were selected as representatives from each quadrant.

| | | **HEMISPHERE** | |
|---|---|---|---|
| | | *Left* | *Right* |
| **LOBE** | *Anterior* | F1, **F3**, F5, Fc1, Fc3, Fc5 | F2, **F4**, F6, Fc2, Fc4, Fc6 |
| | *Posterior* | Cp1, Cp3, Cp5, P1, **P3**, P5 | Cp2, Cp4, Cp6, P2, **P4**, P6 |

manipulation in sentence comprehension. The P200 ERP component to critical words of incongruent sentences was ***less positive*** than that to congruent critical words and the effect was more salient on the left hemisphere. This observation was supported by the significant CONGRUENCY × HEMISPHERE interaction shown both in ANOVA and in regional-averaging ANOVA, $F(1, 19) = 12.99$, $p < 0.01$ and $F(1, 19) = 10.60$, $p < 0.01$, respectively. The main effect of CONGRUENCY was marginal, $p = 0.051$ in ANOVA and $p = 0.032$ in regional-averaging ANOVA. In addition, the LOBE effect neither interacted with CONGRUENCY nor with HEMISPHERE, though regional-averaging ANOVA showed a marginal LOBE × HEMISPHERE interaction, $F(1, 19) = 4.37, p = 0.050$, with the amplitude being the largest in the right posterior region. The results of the ANOVAs were tabulated in Table 5.4 (p. 106).

The summary of the electrophysiological findings regarding the P200 was depicted in Figure 5.5 (p. 107). A general P200 reduction elicited by incongruent sentences was observed and it was modulated by hemisphere, namely, semantic anomalies affected the P200 measured from the left hemisphere most.

This observation was followed by post hoc analysis where for each LOBE-

Table 5.4: Results of the analysis of variance (ANVOA) and regional-averaging ANVOA on the mean voltages in the **P200** time window, 150–200 ms.

| | Factors | | | | | | | $df.$ | $F/t^{\dagger}$ | $p^{\ddagger}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **ANOVA** | | | | | | | | | | |
| | CONGRUENCY | | | | | | | 1,19 | 4.357 | 0.051 |
| | | | LOBE | | | | | 1,19 | 8.314 | 0.010 |
| | | | | | | HEMISPHERE | | 1,19 | 0.549 | 0.468 |
| | CONGRUENCY | × | LOBE | | | | | 1,19 | 1.362 | 0.258 |
| (a) | CONGRUENCY | | | × | | HEMISPHERE | | 1,19 | 12.992 | 0.002 |
| | | | LOBE | × | | HEMISPHERE | | 1,19 | 0.845 | 0.370 |
| | CONGRUENCY | × | LOBE | × | | HEMISPHERE | | 1,19 | <0.001 | 0.985 |
| **Post hoc analysis for (a) on CONGRUENCY** | | | | | | | | | | |
| | in left-anterior quadrant | | | | | | | 19 | 1.933 | 0.068 |
| | in right-anterior quadrant | | | | | | | 19 | 0.404 | 0.690 |
| | in left-posterior quadrant | | | | | | | 19 | 3.646 | 0.002 |
| | in right-posterior quadrant | | | | | | | 19 | 1.571 | 0.133 |
| **Regional-averaging ANOVA** | | | | | | | | | | |
| | CONGRUENCY | | | | | | | 1,19 | 5.325 | 0.032 |
| | | | LOBE | | | | | 1,19 | 2.993 | 0.100 |
| | | | | | | HEMISPHERE | | 1,19 | 2.259 | 0.149 |
| | CONGRUENCY | × | LOBE | | | | | 1,19 | 0.317 | 0.518 |
| (b) | CONGRUENCY | | | × | | HEMISPHERE | | 1,19 | 10.602 | 0.004 |
| | | | LOBE | × | | HEMISPHERE | | 1,19 | 4.374 | 0.050 |
| | CONGRUENCY | × | LOBE | × | | HEMISPHERE | | 1,19 | 0.845 | 0.369 |
| **Post hoc analysis for (b) on CONGRUENCY** | | | | | | | | | | |
| | in left-anterior quadrant | | | | | | | 19 | 2.204 | 0.040 |
| | in right-anterior quadrant | | | | | | | 19 | 1.160 | 0.260 |
| | in left-posterior quadrant | | | | | | | 19 | 3.585 | 0.002 |
| | in right-posterior quadrant | | | | | | | 19 | 1.631 | 0.119 |

† t-value for post hoc analysis using paired sample t-test
‡ uncorrected p-value for post hoc analysis
Red: $p < 0.01$
Blue: $p < 0.05$

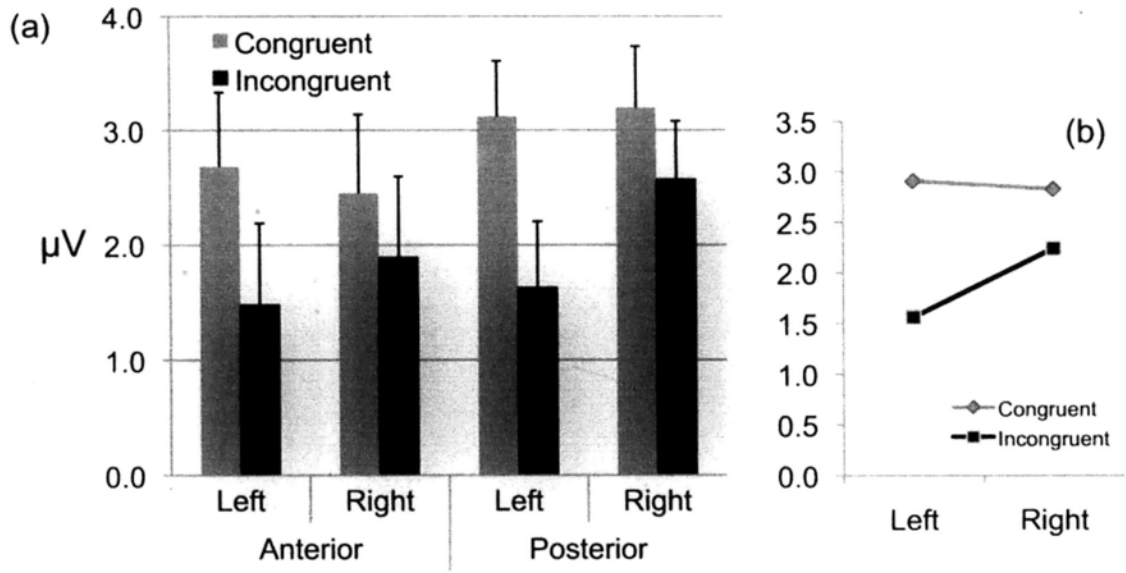



Figure 5.5: Summary of the electrophysiological findings in the P200 region (150 ms to 200 ms). (a) The amplitude, averaged over regions defined in Table 5.3 (p. 105), of the P200 to congruent and incongruent critical words. Each error bar denotes one standard error. (b) The interaction effect of CONGRUENCY × HEMISPHERE revealed by regional-averaging ANOVA.

by-HEMISPHERE quadrant a paired sample t-test was conducted to compare the amplitudes of the P200s elicited by the congruent and incongruent conditions. Post hoc analysis, results also tabulated in Table 5.4 (p. 106), revealed that the effect of CONGRUENCY was significant in the left-posterior quadrant, $t = 3.646$, $p < 0.01$, and $t = 3.585$, $p < 0.01$ (with regional-averaging). In the left-anterior quadrant, however, the effect of CONGRUENCY was marginal, $t = 1.933$, $p = 0.068$, and $t = 2.204, p = 0.040$ (with regional-averaging).

Taken together, the analysis suggested that the P200 effect is lateralised to the left hemisphere with a weak posterior distribution, consistent with visual inspection of the waveforms plotted in Figure 5.3 (p. 102).

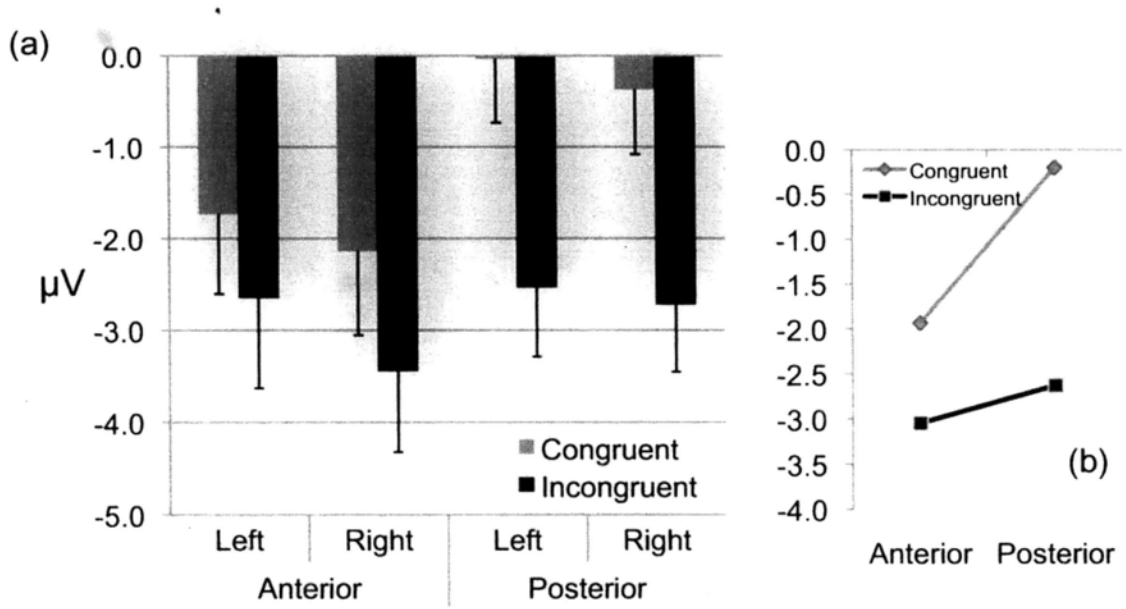



Figure 5.6: Summary of the electrophysiological findings in the N400 region (300 ms to 500 ms). (a) The amplitude, averaged over regions defined in Table 5.3, of the N400 to congruent and incongruent critical words. Each error bar denotes one standard error. (b) The interaction effect of CONGRUENCY × LOBE revealed by regional-averaging ANOVA.

### 5.5.2 The N400 (300–500 ms)

The classic N400 effect as an indicator of anomalies in semantic integration of lexical meaning to the running sentence context was replicated in this study. The N400 ERP component to critical words of incongruent sentences was ***more negative*** that to congruent sentences. This was supported by the significant main effect of CONGRUENCY both in ANOVA and in regional-averaging ANOVA, $F(1, 19) = 15.36$, $p < 0.01$ and $F(1, 19) = 17.91$, $p < 0.01$, respectively. The characteristic scalp distribution of the N400 effect, largest over the posterior sites, was also revealed by the significant CONGRUENCY × LOBE interaction in ANOVA, $F(1, 19) = 18.76$, $p < 0.01$, and in regional-averaging ANOVA, $F(1, 19) = 14.30, p < 0.01$.

The summary of the electrophysiological findings regarding the N400 was depicted in Figure 5.6 (p. 108). The amplitude of the N400 was increased in response to critical words that introduced semantic incongruency. This N400 effect was largest at posterior sites.

This observation was followed by post hoc analysis where for each LOBE-by-HEMISPHERE quadrant a paired sample t-test was conducted to compare the amplitudes of the N400s elicited by the congruent and incongruent conditions. Post hoc analysis, results also tabulated in Table 5.5 (p. 110), revealed that the effect of CONGRUENCY was significant in the left-posterior quadrant, $t = 6.816$, $p < 0.01$, and $t = 5.834$, $p < 0.01$ (with regional-averaging), as well as in the right-posterior quadrant, $t = 5.776$, $p < 0.01$, and $t = 5.823$, $p < 0.01$ (with regional-averaging). In the right-anterior quadrant, however, the effect of CONGRUENCY was marginal $t = 1.906$, $p = 0.072$, and $t = 2.446$, $p = 0.024$ (with regional-averaging). Taken together, the analysis suggested that the N400 effect exhibited a bi-lateral posterior distribution.

### 5.5.3 Summary

A significant P200 reduction associated with semantic incongruency was found. Critical words that were incongruent with the sentence contexts elicited a less positive P200. The interaction effect of CONGRUENCY and HEMISPHERE indicates that the P200 effect was lateralised to the left hemisphere. This suggests an underlying latent component with an unilateral distribution that is sensitive to the mismatch between

Table 5.5: Results of the analysis of variance (ANVOA) and regional-averaging ANVOA on the mean voltages in the **N400** time window, 300–500 ms.

| | Factors | | | | | | $df.$ | $F/t^{\dagger}$ | $p^{\ddagger}$ |
|---|---|---|---|---|---|---|---|---|---|
| **ANOVA** | | | | | | | | | |
| | CONGRUENCY | | | | | | 1,19 | 15.360 | 0.001 |
| | | | LOBE | | | | 1,19 | 5.270 | 0.033 |
| | | | | | | HEMISPHERE | 1,19 | 3.264 | 0.087 |
| (a) | CONGRUENCY | × | LOBE | | | | 1,19 | 18.757 | <0.001 |
| | CONGRUENCY | | | × | | HEMISPHERE | 1,19 | 0.035 | 0.853 |
| | | | LOBE | × | | HEMISPHERE | 1,19 | 0.784 | 0.387 |
| | CONGRUENCY | × | LOBE | × | | HEMISPHERE | 1,19 | 4.816 | 0.041 |
| **Post hoc analysis for (a) on CONGRUENCY** | | | | | | | | | |
| | in left-anterior quadrant | | | | | | 19 | 0.606 | 0.552 |
| | in right-anterior quadrant | | | | | | 19 | 1.906 | 0.072 |
| | in left-posterior quadrant | | | | | | 19 | 6.816 | <0.001 |
| | in right-posterior quadrant | | | | | | 19 | 5.776 | <0.001 |
| **Regional-averaging ANOVA** | | | | | | | | | |
| | CONGRUENCY | | | | | | 1,19 | 17.910 | <0.001 |
| | | | LOBE | | | | 1,19 | 6.640 | 0.018 |
| | | | | | | HEMISPHERE | 1,19 | 4.886 | 0.040 |
| (b) | CONGRUENCY | × | LOBE | | | | 1,19 | 14.292 | 0.001 |
| | CONGRUENCY | | | × | | HEMISPHERE | 1,19 | 0.324 | 0.576 |
| | | | LOBE | × | | HEMISPHERE | 1,19 | 1.983 | 0.175 |
| | CONGRUENCY | × | LOBE | × | | HEMISPHERE | 1,19 | 3.292 | 0.085 |
| **Post hoc analysis for (b) on CONGRUENCY** | | | | | | | | | |
| | in left-anterior quadrant | | | | | | 19 | 1.809 | 0.086 |
| | in right-anterior quadrant | | | | | | 19 | 2.446 | 0.024 |
| | in left-posterior quadrant | | | | | | 19 | 5.834 | <0.001 |
| | in right-posterior quadrant | | | | | | 19 | 5.823 | <0.001 |

† t-value for post hoc analysis using paired sample t-test

‡ uncorrected p-value for post hoc analysis

Red: $p < 0.01$

Blue: $p < 0.05$

the sentence context and the lexical meaning. The classic N400 effect, enhanced negativities associated with semantic incongruency, was also replicated.

## 5.6 Discussion

In this study, semantic anomalies were introduced by exchanging critical words between pairs of congruent sentences. I designed my test materials such that the exact same set of critical words were used to elicit both the congruent and the incongruent conditions in order to balance factors like the frequency of usage, the visual complexity of the critical words (e.g. number of strokes) as well as their lexical meanings. I observed that the brain's early sensitivity to the semantic manipulation was manifested as a change in the amplitude of ERP that peaked in the time window of 150 ms to 200 ms after stimulus onsets, the P200, indicating an early contextual influence during sentence comprehension.

It has been widely accepted that the N400 effect is interpreted as an increased work load for contextual integration (Hagoort et al., 2004; Hald, Bastiaansen & Hagoort, 2006; Kutas & Federmeier, 2000). Extending such an interpretation to my findings in the early time window may appear as contradictory because the P200 was *reduced* in response to semantic incongruency. My explanation to this apparent contradiction is that a peak in a grand average waveform is not equal to just a single latent component (Luck, 2005). It is thus possible that within the observed P200 peak, there are latent components whose negativities were enhanced by the experimental manipulation. The

observation of a negative-going deflection, albeit small, in the incongruent condition within the P200 region, most observable for electrodes P3 and Pz (Figure 5.3 in p. 102), is consistent with this explanation.

From recent studies on brain functions in language processing during the early time window, the picture that is emerging is much more complex. A P200 and an N200[7] effect was reported by Federmeier et al. (2005) and Ruz & Nobre (2008), respectively. The authors demonstrated that ERP components at this time region were modulated by attention. They both made a conclusion that the effect on early ERP components indicates a top-down enhancement on the initial stages of visual perception. The visual roles of P200 were also elaborated as visual feature detection (Luck, 2005), graphic processing of Chinese characters (Liu et al., 2003; Perfetti, Liu, Fiez, Nelson, Bolger & Tan, 2007) as well as grapheme-to-phoneme conversion (Lee, Tsai, Chiu, Tzeng & Hung, 2006; Proverbio, Vecchi & Zani, 2004). These studies collectively argued for a *perceptual expectation* account of the P200 effect whereby an enhanced P200 is supposed to reflect the mismatch between the visual forms of the expected words and the actually presented words. In the other words, the P200 effect is suggested to be the visual analogue of the phonological mismatch negativities (Connolly & Philips, 1994).

My data is compatible with either the perceptual expectation account or a *semantic expectation* account, which suggests that the P200 reflects an outright semantic incongruency. Although the experiment reported in this study cannot provide direct

[7] A P200 effect was reported in Federmeier, Mai & Kutas (2005) using average-mastoid as the re-reference montage; on the other hand, an N200 effect was reported in Ruz & Nobre (2008) using average-reference. I consider the inverse polarity mainly the consequence of the choice of reference.

evidence for or against either account, a clearer picture emerges when we take into consideration the similarity in polarity of the ERP effects across studies. The auditory N200 and N270 effects in the study of van den Brink et al. (2001) and Connolly & Philips (1994), respectively, both were *enhanced negativities* in response to word-context incongruency. The P200 effect of my study, and of Meng et al. (2007), manifested as a *reduced positivity*, i.e. towards the negative side, to words that were incongruent with the contexts. This consistent polarity of effects to incongruency could be explained by the presence of an underlying latent component whose negativity was enhanced by the mismatch with contextual expectations. Critically, this latent component appear to be modality independent, a situation that is less probable under a pure perceptual expectation account of the P200 effect. I therefore argue that an early semantic integration is already taking place during the time window indexed by the P200.

My results suggested that at the latest of 150 ms to 200 ms after stimulus onset, the brain is already processing the lexical meaning of a word under the influence of the sentence context in which the word appears. I consider such findings to be direct evidence challenging the serial, syntax-first sentence processing model (Friederici, 2002), in which auditory sentence comprehension is divided into temporally ordered phases of processing:

**Phase 0** Identification of phonemes, as indexed by N100 (at around 100 ms after stimulus onset);

**Phase 1** Identification of syntactic categories, as indexed by early left-anterior negativity LAN (at a time window of 150–200 ms);

**Phase 2** Morphosyntactic and lexical-semantic processing, as indexed by LAN and N400, respectively (at a time window of 300–500 ms);

**Phase 3** Integration of the information from different levels of analysis, as indexed by P600 (at around 600 ms)

Friederici (2002) characterised Phase 1 as the period in which *"initial syntactic structure is formed on the basis of information about word category"* and Phase 2 as the period in which *"lexical-semantic and morphosyntactic processes take place with the goal of thematic role assignment"*. This model also assumes that early syntactic processing is independent of lexical semantics and the interaction of syntax and semantics only happens during a later time frame, (Phase 2 or Phase 3).

The P200 effect reported in this study is therefore contradictory to the predictions of a serial, syntax-first model because under Friederici's framework, the lexical meaning of a word has not even been identified nor influence brain processes within this early time frame.

In conclusion, I propose a scenario to account for the "double-take" effect, the P200 and N400, on ERP to semantic incongruency. I conjecture that during reading, the brain is constantly ready to make use of whatever information is available in order to meet the requirement for real-time comprehension. In other words, semantic

integration has constantly been in action during in-context word recognition. It appears as discrete stages in the electrophysiological manifestations due to the time lag between the availability of progressively more elaborate information. In the case of reading, initial lexical access is triggered by the product of (ortho-)graphic analysis of the characters, as reflected by the P200 response. Though the lexical semantics has not yet been fully revealed, the brain has already started attempting to make sense of the semantic features available at this level of analysis, as reflected by the P200 effect on congruency. This effect might be more salient in Chinese readers than in readers of European languages. Possibly because written Chinese has a stronger association between graphics and semantics than languages with alphabetical writing systems (Wang, 1973). The morphosyllabic characteristics of the Chinese writing systems make it possible for experienced readers to extract the meanings of a word directly from its written form (Perfetti & Tan, 1998).

## 5.7 Conclusion of and summary of Chapter 5

An event-related electrical potential (ERP) study was conducted to investigate early neurophysiological response to contextual influences in reading. Both the classic N400 component and an early ERP component, the P200, were found to be sensitive to the experimental manipulations on congruency. I employed a design of the test materials such that the exact same set of words were used to elicit both the congruent and the incongruent condition. This way I kept to a minimum the contamination of experimental effects due to stimulus variability enabling us to reveal the P200 effect

that might have gone unnoticed otherwise. A reduction in amplitude of the P200 to incongruent critical words was observed and it was lateralised to the left hemisphere. My results provide evidence that within the first 200 ms the brain has already started to process the lexical meanings of a word and this process interacts with the information carried by the sentence context. Such an early brain signature to semantic processing poses a challenge to the serial, syntax-first model of language processing. Implications for theories of Chinese reading was also discussed.

# Chapter 6

# General Discussion

## 6.1 Combinatorial productivity, why should we care?

Marcus (1998) once illustrated the limitation of connectionist modelling with a language learning scenario consisted of the following seven sentences:

| |
|---|
| "the bee sniffs the rose" |
| "the bee sniffs the lily" |
| "the bee sniffs the tulip" |
| "the bee sniffs the lilac" |
| "a rose is a rose" |
| "a lily is a lily" |
| "a tulip is a tulip" |

He trained simple recurrent networks with two sentence frames: "the bee sniffs the X" and "a Y is a Y". The words "rose", "lily" and "tulip" appeared in both sentence

frames but "lilac" occurred only in the first one.[1] Marcus showed that, after training, SRNs failed to make the prediction that "lilac" is a grammatical continuation of the testing set sentence "an lilac is an ..." and therefore demonstrated that SRNs cannot generalise from the training data. Marcus argued that connectionist networks lack an abstract representation that "rose", "lily", "tulip" and "lilac" are all nouns of this kind and hence fails to generalise the relationship to the sentence frame "a lilac is a lilac" "*in the way humans do*".

I do not doubt the simulation results of Marcus (1998), rather I take this as an common misconception of modelling learning and generalisation. Generalisation is a transfer of knowledge between two domains. In the case of modelling with artificial neural networks, between the training set and the testing set; in the case of natural language acquisition, between the utterances a child *has heard* to the utterances that he or she *will* eventually hear.

This transfer of knowledge does not come out of the blue, it requires a similarity to be established in the learners' mind between the two domains. The required similarity is often a functional one, for instance, in Marcus's illustration, that "lilac" is similar to "rose", "lily" and "tulip" in a sense that they are all objects that bees sniff.

The illustration of Marcus's (1998) can be invalidated simply by considering training the networks with "the bee sniffs the Y" and "a X is a X"[2] instead. In this scenario, if an SRN made the prediction that "lilac" is a grammatical continuation

[1] Hence, X={rose, lily, tulip, lilac} and Y={rose, lily, tulip}
[2] Still, X={rose, lily, tulip, lilac} and Y={rose, lily, tulip}

of "`the bee sniffs the...` ", we could as well conclude that it is an error of *over generalisation*.

What I want to illustrate with the two scenarios is that quite often the criticism that connectionist networks fail to exhibit human-like ability to generalise is often poorly formulated when it comes to concrete terms as required by computational simulations. In the case of the simulation of Marcus (1998), a rich enough linguistic environment was not established for the networks to construct the "abstract representation" that Marcus argued to be the simple solution to the problem of generalisation.

This defence of connectionist modelling echoes the view of Rohde & Plaut (1999):

> "*... in developing a model of how people behave with particular items in particular contexts, it is rarely adequate to train a network on only those items in those contexts. Rather, a network would be expected to generalize the way people do only if its training environment adequately approximated the full range of relevant surface and functional similarities that people experience in the domain.*"
>
> Rohde & Plaut (1999, p. 298)

The issue of **combinatorial productivity** first raised as a challenge to connectionist modelling by van der Velde et al. (2004), however, was grounded with a well formulated simulation environment as well as empirical findings [3] from child language acquisition.

[3]Though none was not cited in van der Velde et al. (2004)

Recall that in the simulations reported in Chapter 2 and Chapter 4, the SRNs were trained with sentences, such as, among many others:

---

"the dog chased the cat"

"the boy saw the girl"

---

They were tested if they could generalise to sentences that were composed of novel combinations of lexical items, such as, among many others:

---

"the dog saw the cat"

"the boy chased the girl"

---

I have argued that success in generalisation of this sort depends on networks' ability to induce that "cat" and "girl" share a similarity that they are both preceded by a verb and it has a certain syntactic relationship with a sentence initial noun. Given that there are sufficient instances of such, the networks are therefore expected to generalise and indeed my simulations have provided evidence that networks do so.

On the empirical side, one might ask if combinatorial productivity is a real phenomenon in natural language acquisition. An intuitive confirmation is that normal adults would have no problem in answering the question of "who has done what to whom" to any arbitrary sentences that adhere to a Noun-Verb-Noun sentence frame and even to sentences that contain words that he or she has never encountered before. For example if asked a comprehension question "whom did the prince kiss?" concerning

Table 6.1: Examples of sentences used in Valian et al. (2006). The numbers in the right column are the token counts I gathered through collocation analysis on the British National Corpus (BNC) via *Sketch Engine*†. Each number corresponds to the number of sentences, in the entire BNC, in which the nouns (e.g. game) were to the right of the verbs (e.g. play), within a less than 15 words separation. There were about 97 million words in the BNC.

| **Sentence** | **Collocation in BNC** |
|---|---|
| The frog is ***playing*** a ***game*** (***drum***). | 2,405 (73) |
| The horse ***sings*** a ***song*** (***story***). | 641 (30) |
| The clown is ***driving*** a ***car*** (***boat***). | 1,239 (32) |
| The cowboy is ***cooking*** our ***dinner*** (***crackers***). | 921 (0) |

†Kilgarriff et al. (2004), http://www.sketchengine.co.uk

the sentence "`the prince kissed etihwwons`", one could still give the correct answer "etihwwons" even if it is the first time he or she encounter such a word.

Combinatorial productivity also manifests itself in child language acquisition. In the study conducted by Valian, Prasada & Scarpa (2006), children aged 21–35 months were tested with a comprehension task on spoken sentences. These sentences contained direct objects that were either "predictable" or "unpredictable" with respect to the verbs. Examples of the experimental sentences used in Valian et al. (2006) were given in Table 6.1 (p. 121) where the words in parenthesises were the unpredictable direct objects with respect to the verbs.

After hearing the sentences, the children were able to perform correctly a comprehension task upon requests such as "*Show me the animal that played a game*"[4] to the first sentence listed in Table 6.1 (p. 121). Interestingly, children performed equally

[4]Children responded to each sentence by placing a sticker on one of the two pictures shown to them, one of which was the correct subject to the sentence they had spoken to by the experimenters.

well to predictable and unpredictable sentences, with an overall 70% correct sticker placement.

The unpredictable sentences used in Valian et al. (2006) were analogous to the testing set sentences I have used to evaluate SRNs' potential to exhibit combinatorial productivity in the simulation reported in Chapter 2 and Chapter 4. Unpredictable sentences in Valian et al. (2006) such as "the cowboy is cooking our crackers" could have been novel to the children in a sense that the word "crackers" rarely collocates with the word "cook", cf. Table 6.1 (p. 121).

To conclude, I agree with van der Velde et al. (2004) that combinatorial productivity is one of the essential feature of natural language. Human language learners need to possess the ability to generalise combinatorially in order to be able to deal with the ever growing language. This essential learnability problem has not been addressed before though it has a close relationship with the long-standing notion of compositionality (Fodor & Pylyshyn, 1988).

## 6.2 Time course of reading revealed by EEGs, what's next?

In the second part of this thesis, I reported my study that made use of a high temporal resolution brain imaging technique, electroencephalography (EEG), to investigate the time course of brain activities during the task of reading for comprehension. The important finding of my study was the observation that at the ***latest*** of 200 ms the

brain has already been processing the meanings of a word under the influence of the sentence context. This was supported by the observed semantic modulation of the P200 ERP component, positivity at around 150–200 ms after stimulus onset, where words that were incongruent with the sentence context elicited a smaller, less positive, event related potential (ERP) compared to the congruent words.

This search of the earliest brain response to semantics has profound implications for theories of language processing in general. As the presence of an early ERP component to meaning not only brings a new member to the family of major language-related ERPs, but also acts as direct evidence against the serial, syntax-first model of language processing, as proposed by Friederici (2002) that auditory sentence comprehension is divided into temporally ordered phases of processing:

(i) Identification of phonemes, as indexed by N100 (at around 100 ms after stimulus onset);

(ii) Identification of syntactic categories, as indexed by early left-anterior negativity LAN (at a time window of 150–200 ms);

(iii) Morphosyntactic and lexical-semantic processing, as indexed by LAN and N400, respectively (at a time window of 300–500 ms);

(iv) Integration of the information from different levels of analysis, as indexed by P600 (at around 600 ms)

Finding early brain response to semantics also has it impacts on theories of Chinese word reading. Like most orthographic systems, the Chinese characters code for the sounds of the words that in turn mediate the identification of the lexical meanings (Tzeng & Wang, 1983). However, unlike alphabetic writing systems, such as English, the Chinese writing systems do so at a *syllable level* rendering grapheme-to-phoneme mapping impossible (Perfetti & Tan, 1998). For example, no parts of the character "big" (pronounced *da*, Figure 6.1 (a) in p. 125) denote either the initial *d* or the final *a*. Also, the mapping of the graphic forms to sounds is not systematic. Though the characters "big" and "too" (Figure 6.1 (a) and (b), respectively) both "look alike" and "sound alike", there are abundance of counterexamples like "dog" and "fire" (Figure 6.1 (c) and (d), respectively) that "look alike" but do not share any phonetic segmental features with one another. Meanwhile written Chinese has a stronger association between graphics and semantics (Wang, 1973) as illustrated by the character "horse" in Figure 6.1 (e). The abstract shape of the animal horse resembles the shape of the character.

On one hand, the above morphosyllabic characteristics of the Chinese writing systems make it possible for experienced readers to extract the meanings of a word directly from its written form without first activating the phonology of the word (Perfetti & Tan, 1998). On the other hand, the contribution of phonology to a better memory, via *speech recoding*, of written materials (Tzeng, Hung & Wang, 1977) suggests that phonological activation may be of equal importance for reading and comprehension.

The issue of the time course of brain activities in reading Chinese words could be put

(a) 大 "big", as in *a big mouse*
JP: *daai*$^{6}$ PY: *da*$^{4}$

(b) 太 "too", as in *too good to be true*
JP: *taai*$^{3}$ PY: *tai*$^{4}$

(c) 犬 "dog", as in *the dog barks*
JP: *hyun*$^{2}$ PY: *quan*$^{3}$

(d) 火 "fire", as in *the fire hurts me*
JP: *fo*$^{2}$ PY: *huo*$^{3}$

(e) 馬 "horse", as in *riding a horse*
JP: *maa*$^{5}$ PY: *ma*$^{3}$

Figure 6.1: Samples of traditional Chinese characters showing the distinctive features of the Chinese writing systems. The corresponding meanings and pronunciations are shown in the first and the second line, respectively. Romanisation of the Cantonese and the Mandarin pronunciation follows the Jyutping (JP) and Pingyin (PY) system, respectively. The superscripts denote the tone categories of the syllables.

as a simple, yet controversial, empirical question: Upon seeing a word, is phonological processing activated before semantic processing?

My findings about the semantic role of the P200 effect has provided some support to the view that semantic processing does not need to wait until phonological activation. And semantic processing of word is started very early, within one fifth of a second. However, my ERP data, as of today, cannot provide an definite answer because a time stamp of the earliest ERP component to phonological processing was not identifiable in the study that I have conducted. This will be a next step of my research on language and the brain.

## 6.3 Conclusion

In closing, I have reported my study towards the understanding of the neurophysiological bases of language. The approach I haven take with computation modelling as reported in the first part of the thesis is more inclined to answering a theoretical concern for the computational adequacy of connectionist networks. This was first raised by van der Velde et al. (2004) in which they argued that simple recurrent networks fail to generalise to meet the combinatorial productivity of natural language. I argued otherwise and demonstrated cases in which networks could do so. I illustrated that, through the analyses of networks' internal representation, networks that were more successful in generalisation developed categories according to syntactic positions.

The second part of the thesis reflects the cognitive side of my work. An event-related brain potential study was conducted to investigate how a living brain functions during the course of reading for comprehension. The major finding of this brain waves study was that an early ERP component, the P200, was found to be sensitive to the experimental manipulations on semantic congruency. This is a piece of evidence suggesting that within one fifth of a second the brain has already started to process the meanings of a word.

# Appendix A

# Plots of SRNs' output layer activations, SRN-SIM1

**Figure A.1 (p. 128)** The time course of output layer activations of the twenty **single hidden layer** SRNs with 40 hidden layer neurons during the processing of (a) a training set right-branching sentence and (b) a training set centre-embedding sentence. The networks were trained with the procedure described in Chapter 2.

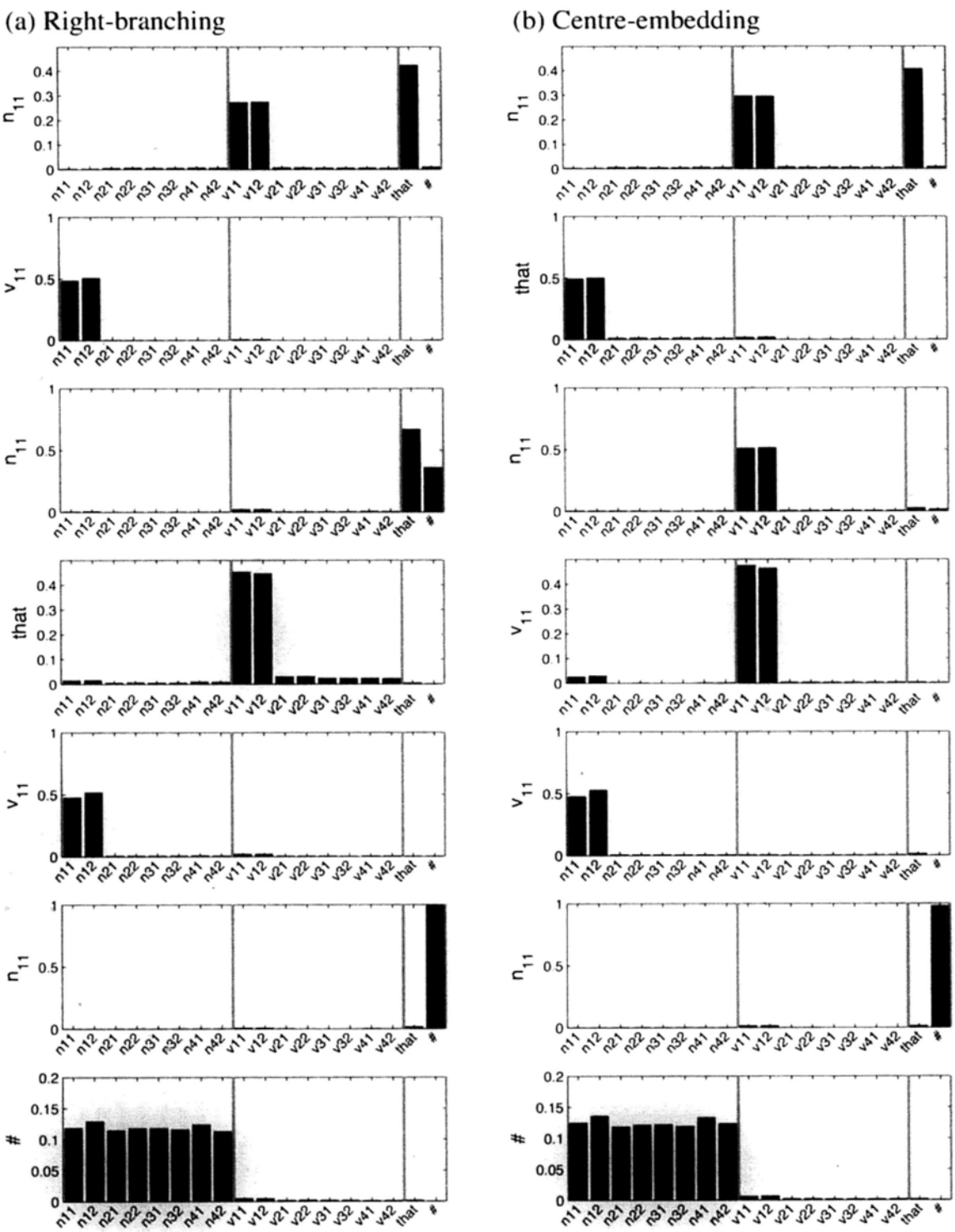



Figure A.1: The time course of output layer activations of the twenty **single hidden layer** SRNs

# Appendix B

# Plots of 2-hidden-layer-SRNs' hidden layer activations, SRN-SIM2

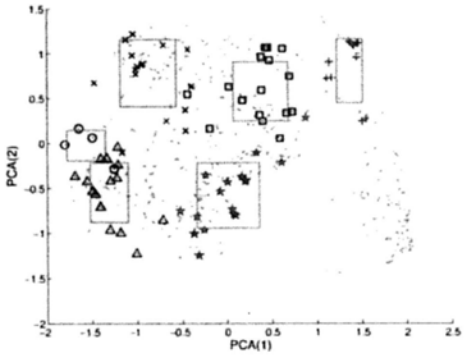
**Table B.1 (p. 130)** Hidden layer activations during the processing of $M = 4$ **right-branching sentences** obtained from SRNs that were less successful in generalisation (network #5 and #10) and SRNs that were more successful in generalisation (network #8 and #12), see Section 4.7.2 (p. 74) for details.

**Table B.2 (p. 131)** Hidden layer activations during the processing of $M = 4$ **centre-embedding sentences** obtained from SRNs that were less successful in generalisation (network #5 and #10) and SRNs that were more successful in generalisation (network #8 and #12), see Section 4.7.2 (p. 74) for details.

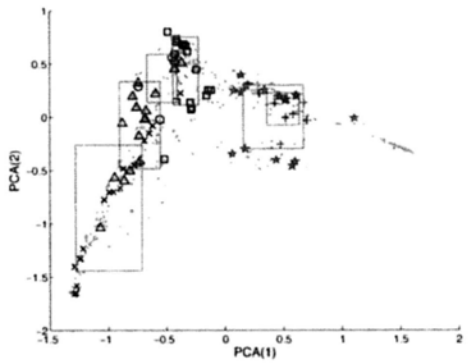
Table B.1: Right-branching sentences

Network #5, **high** GPE attained

1st hidden layer

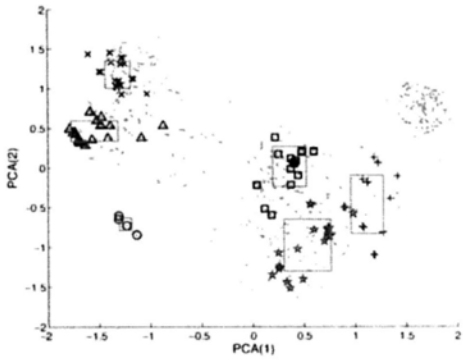



2nd hidden layer

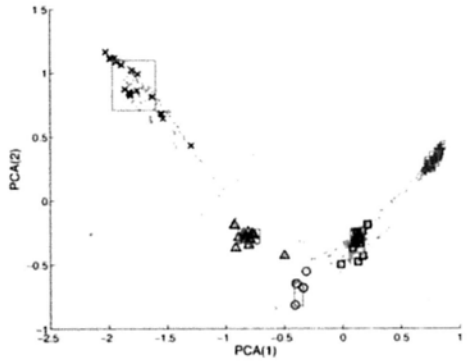



Network #8, **low** GPE attained

1st hidden layer

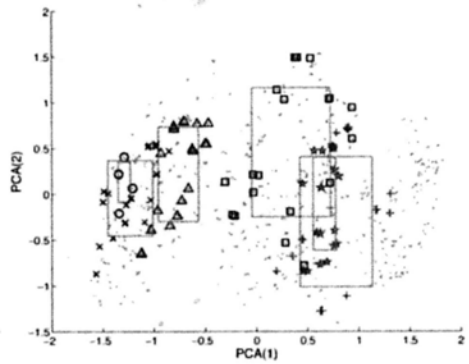



2nd hidden layer

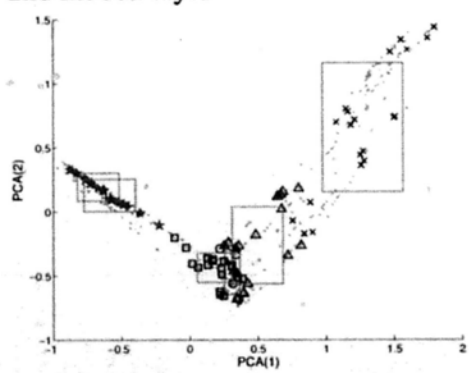



Network #10, **high** GPE attained

1st hidden layer

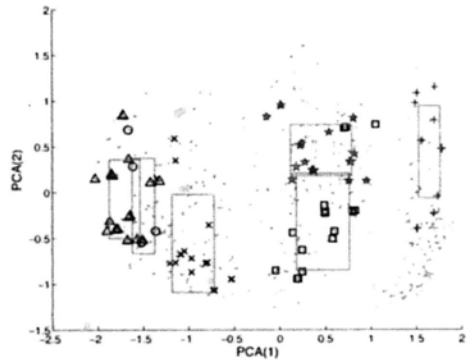



2nd hidden layer




Network #12, **low** GPE attained

1st hidden layer




2nd hidden layer

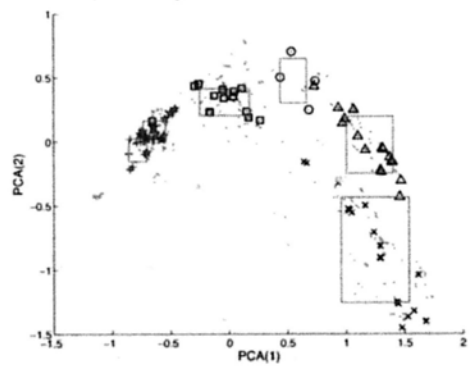
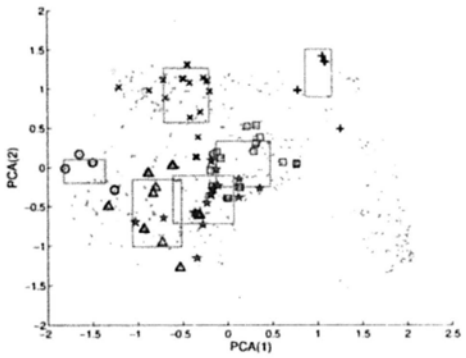


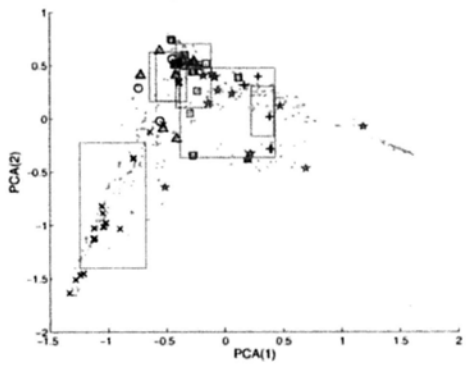
Table B.2: Centre-embedding sentences

Network #5, **high** GPE attained

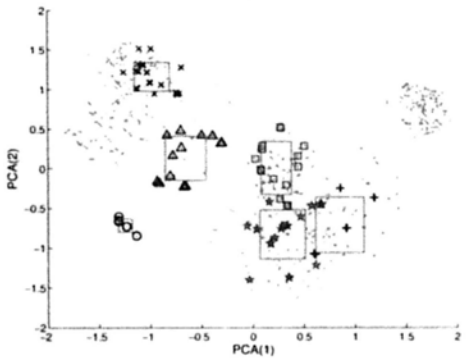
1st hidden layer

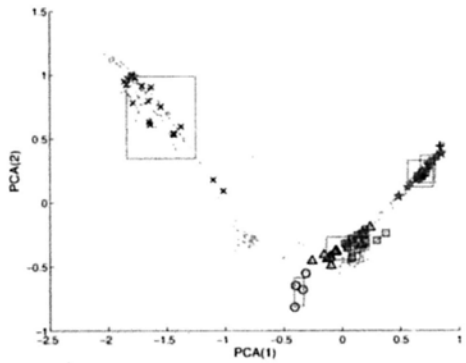



2nd hidden layer



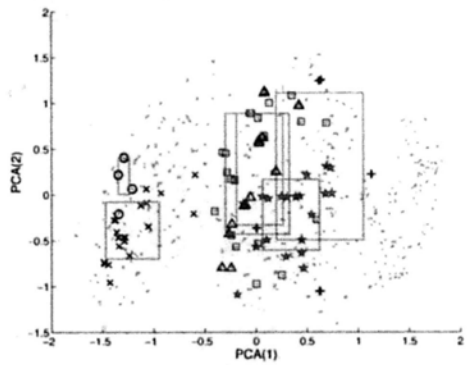

Network #8, **low** GPE attained



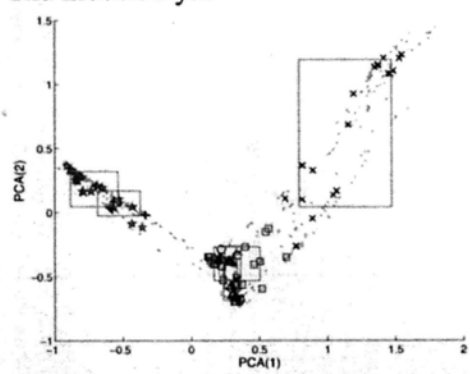




Network #10, **high** GPE attained

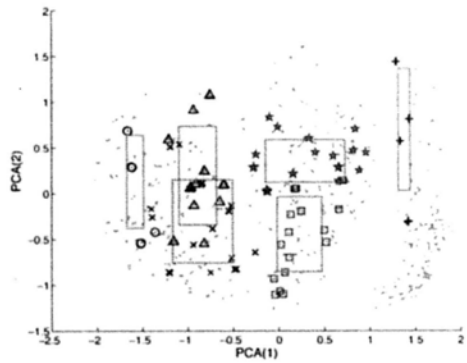
1st hidden layer

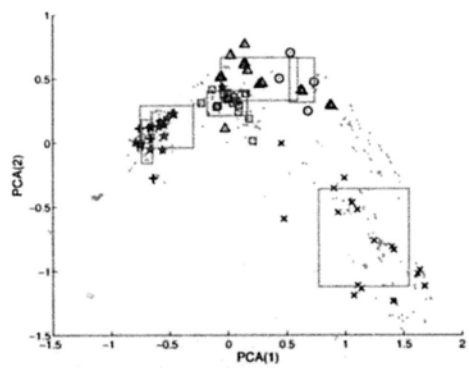



2nd hidden layer

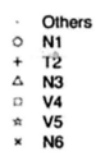



Network #12, **low** GPE attained







# Appendix C

# Materials used in the ERP study reported in Chapter 5

**Table C.1 (p. 133)** The 58 experimental sentences. For each sentence, the Chinese version is shown in the first line. The second line gives the corresponding English gloss. The third line gives the translation in English. Words in brackets were used to replace the congruent critical words, words that immediately preceded the bracketed words, to introduce the incongruent version of each of the congruent sentences. The symbol "|" denotes the segmentation of the sentences into words. See Section 5.3.2 (p. 90) for details.

**Table C.2 (p. 140)** Handedness questionnaire used in the study, adopted from Snyder & Harris (1993) and Tan et al. (2000).

**Table C.3 (p. 141)** Instruction sheet delivered to participants before the start of each EEG session.

Table C.1: The 58 experimental sentences, $S_x$ for $x = \{1, 2, 3...58\}$

| $S_x$ | Notation | The sentence |
|---|---|---|
| 1 | Chinese | 藏有 \| 偽造 \| 證件 \| (機會) \| 是 \| 違法的 |
| | Gloss | in-possession \| forged \| credentials \| (risk) \| is \| an-offence |
| | English | It is an offence in law to be in possession of forged credentials |
| 2 | Chinese | 經常 \| 接觸 \| 二手煙 \| 患癌的 \| 機會 \| (證件) \| 較 \| 其他人 \| 高 |
| | Gloss | often \| exposed-to \| second-hand-smoking \| getting-cancer \| risk \| (credentials) \| more \| others \| high |
| | English | There is an increase risk of cancer of people who are often exposed to second-hand smoking |
| 3 | Chinese | 該名 \| 男子 \| 頭部 \| 嚴重 \| 受傷 \| 被 \| 送往 \| 醫院 \| (病毒) \| 救治 |
| | Gloss | the \| man \| head \| serious \| injury \| was \| taken-to \| hospital \| (virus) \| treatment |
| | English | Sustaining serious head injury, the man was taken to the hospital for medical treatment |
| 4 | Chinese | 醫學界 \| 對 \| 這種 \| 前所 \| 未見的 \| 病毒 \| (醫院) \| 所知 \| 不多 |
| | Gloss | medical-profession \| of \| this-kind \| ever \| unseen \| virus \| (hospital) \| knowledge-of \| not-much |
| | English | The medical profession had little knowledge of this new virus |
| 5 | Chinese | 報告 \| 定下了 \| 多項 \| 改善 \| 空氣 \| 質素 \| (證書) \| 的 \| 措施 |
| | Gloss | report \| has-set \| several \| improvement \| air \| quality \| (certificate) \| AS \| policies |
| | English | The report set several policies for improving air quality |
| 6 | Chinese | 出席率 \| 較高的 \| 參加者 \| 將獲 \| 頒發 \| 證書 \| (質素) |
| | Gloss | attendance-rate \| higher \| participant \| will-gain \| award \| certificate \| (quality) |
| | English | Participants with high attendance rate will be awarded a certificate |
| 7 | Chinese | 內地 \| 與 \| 香港 \| 有 \| 互利的 \| 平等 \| 關係 \| (兒童) |
| | Gloss | mainland \| with \| hong-kong \| have \| mutually-complementary \| equal \| relationship \| (children) |
| | English | The relationship between the Mainland and Hong Kong is equal and mutually complementary |
| 8 | Chinese | 在 \| 擠迫的 \| 地方 \| 要 \| 自律 \| 並要 \| 對 \| 老人 \| 及 \| 兒童 \| (關係) \| 禮讓 |
| | Gloss | in \| crowded \| place \| should \| self-restraint \| and-should \| to \| elderly \| and \| children \| (relationship) \| courteous |
| | English | The public are also urged to exercise restraint and to be courteous to the elderly and children in crowded places |
| 9 | Chinese | 今日 \| 不少 \| 邊緣 \| 青年 \| (球鞋) \| 都是 \| 來自 \| 問題 \| 家庭 |
| | Gloss | nowadays \| many \| at-risk \| young-people \| (sports-shoes) \| are \| from \| problem \| families |

Continuation of Table C.1

| | | |
|---|---|---|
| | English | Nowadays many young people at risk come from families experiencing problems |
| 10 | Chinese | 機場 \| 海關 \| 檢獲 \| 冒牌 \| 球鞋 \| (青年) |
| | Gloss | airport \| customs \| seize \| counterfeit \| sports-shoes \| (young-people) |
| | English | Airport Customs seized counterfeit sports shoes |
| 11 | Chinese | 不少人 \| 以為 \| 只要 \| 利用 \| 英語 \| 教學 \| 英語 \| 能力 \| (罰款) \| 自然會 \| 好起來 |
| | Gloss | many-people \| have-misconception \| only \| make-use-of \| English \| to-teach \| English \| ability \| (fine) \| naturally \| get-better |
| | English | There is a common misconception in the community that English proficiency will come as a natural result of using English as the medium of instruction |
| 12 | Chinese | 一名 \| 僱主 \| 因 \| 拖欠 \| 工資 \| 被判 \| 罰款 \| (能力) |
| | Gloss | an \| employer \| because-of \| behind-on-payment \| wage \| be-convicted \| fine \| (ability) |
| | English | An employer was fined for wage offences |
| 13 | Chinese | 面對 \| 知識型 \| 經濟 \| 我們 \| 會 \| 大力 \| 投資 \| 教育 \| (市民) |
| | Gloss | to-prepare \| knowledge-based \| economy \| we \| will \| heavily \| invest \| education \| (citizen) |
| | English | To prepare for the knowledge-based economy we will invest heavily in education |
| 14 | Chinese | 展覽 \| 舉行 \| 期間 \| 吸引 \| 大批 \| 市民 \| (教育) \| 參觀 |
| | Gloss | exhibition \| conduct \| duration \| attract \| many \| citizen \| (education) \| visit |
| | English | The exhibition was well received as it attracted many citizens to visit |
| 15 | Chinese | 政府 \| 透過 \| 科技 \| 提高 \| 道路 \| 使用的 \| 效率 \| (批評) |
| | Gloss | government \| through \| technology \| maximize \| road \| usage \| efficiency \| (criticism) |
| | English | The government try to maximize road use efficiency with the use of advanced technology |
| 16 | Chinese | 局長 \| 接受 \| 報告 \| 裏面的 \| 意見 \| 和 \| 批評 \| (效率) |
| | Gloss | principal-official \| accept \| report \| in-the \| suggestion \| and \| criticism \| (efficiency) |
| | English | The Principal Official accepted the suggestions and criticism in the reported |
| 17 | Chinese | 入境處 \| 扣查 \| 一名 \| 懷疑 \| 僱用 \| 黑工的 \| 僱主 \| (程度) |
| | Gloss | immigration-department \| arrest \| a \| suspected \| hire \| illegal-worker \| employer \| (level) |
| | English | The Immigration Department arrested an employer suspected of hiring illegal workers |
| 18 | Chinese | 失業 \| 人士 \| 中 \| 半數 \| 只有 \| 初中 \| 以下的 \| 教育 \| 程度 \| (僱主) |
| | Gloss | unemployed \| people \| among \| half \| only-have \| junior-secondary \| below \| education \| level \| (employer) |

Continuation of Table C.1

| | | |
|---|---|---|
| | English | Of the unemployed, half are educated to junior secondary level or below |
| 19 | Chinese | 重陽節 \| 前後 \| 是 \| 發生 \| 山火 \| (藥物) \| 的 \| 高危 \| 季節 |
| | Gloss | Chung-Yeung-Festival \| around \| is \| to-happen \| hill-fire \| (drugs) \| AS \| high-risk \| season |
| | English | The time of the year around Chung-Yeung Festival is a high-risk period for hill fires |
| 20 | Chinese | 青少年 \| 濫用 \| 藥物 \| (山火) \| 的 \| 情況 \| 有所 \| 增加 |
| | Gloss | Youngster \| to-abuse \| drug \| (hill-fire) \| AS \| circumstances \| there exists \| increase |
| | English | An increase was observed in the number of young drug abusers |
| 21 | Chinese | 他們 \| 出現 \| 肺炎 \| 病徵 \| 被 \| 送往 \| 醫院 \| 接受 \| 治療 \| (熱線) |
| | Gloss | they \| develope \| pneumonia \| symptom \| is \| admitted \| hospital \| receive \| treatment \| (hotline) |
| | English | they have developed symptoms of pneumonia and were admitted to hospital for treatment |
| 22 | Chinese | 遇到 \| 因 \| 賭博 \| 引起的 \| 問題 \| 可以 \| 致電 \| 熱線 \| (治療) \| 尋求 \| 協助 |
| | Gloss | facing \| due-to \| gambling \| caused-by \| problem \| can \| call \| hotline \| (treatment) \| seek \| help |
| | English | Anyone with problems arising from their gambling behaviour could seek help through the hotline |
| 23 | Chinese | 我們 \| 希望 \| 通過 \| 這批 \| 質素 \| 高的 \| 電影 \| (負擔) \| 引起 \| 觀眾 \| 對 \| 電影的 \| 興趣 |
| | Gloss | we \| hope \| through \| these \| quality \| high \| film \| (burden) \| arouse \| viewer \| to \| film \| interest |
| | English | We hope that these high-quality films will arouse the public's interest in films |
| 24 | Chinese | 居屋 \| 有助 \| 紓緩 \| 政府的 \| 財政 \| 負擔 \| (電影) |
| | Gloss | home-ownership-scheme \| helpful \| alleviate \| of-government \| financial \| burden \| (film) |
| | English | Home Ownership Scheme is helpful to alleviating the financial burden on the Government |
| 25 | Chinese | 財政司 \| 司長 \| 澄清 \| 關於 \| 聯繫 \| 匯率 \| 制度的 \| 謠傳 \| (車輛) |
| | Gloss | financial \| secretary \| dismiss \| about \| linked \| exchange-rate \| of-system \| rumour \| (vehicle) |
| | English | Financial Secretary dismisses rumours about changes in the Linked Exchange Rate System |
| 26 | Chinese | 所有 \| 違法 \| 貨品 \| 及 \| 車輛 \| (謠傳) \| 均可 \| 予以 \| 沒收 |
| | Gloss | any \| an-offence \| articles \| and \| vehicle \| (rumour) \| liable \| will-be \| forfeit |
| | English | Any articles and vehicles found in connection with the commission of an offence are liable to forfeiture |

Continuation of Table C.1

| | | |
|---|---|---|
| 27 | Chinese | 展覽場 \| 裝有 \| 巨型的 \| 海底 \| 世界 \| 噴畫 \| (損毀) |
| | Gloss | exhibition-area \| installed-with \| gigantic \| underwater \| world \| inkjet-print \| (damage) |
| | English | A gigantic inkjet prints showing the underwater world had been installed in the exhibition area |
| 28 | Chinese | 六級 \| 以上的 \| 地震 \| 可以 \| 造成 \| 災難性 \| 損毀 \| (噴畫) |
| | Gloss | magnitude-six \| above \| earthquake \| can \| cause \| catastrophic \| damage \| (inkjet-print) |
| | English | Earthquakes of magnitude six are capable of causing catastrophic damage |
| 29 | Chinese | 康文署 \| 提醒 \| 市民 \| 今晚 \| 九時 \| 有 \| 煙花 \| (支柱) \| 匯演 |
| | Gloss | cultural-services-department \| remind \| public \| tonight \| 9-pm \| have \| fireworks \| (pillar) \| display |
| | English | The Leisure and Cultural Services Department reminds the public of the fireworks display tonight at 9 pm |
| 30 | Chinese | 中小型 \| 企業 \| 是 \| 香港 \| 經濟的 \| 支柱 \| (煙花) |
| | Gloss | small-medium-sized \| enterprise \| is \| hong-knog \| of-economy \| pillar \| (fireworks) |
| | English | Small and medium enterprises had always been a pillar of Hong Kong's economy |
| 31 | Chinese | 環保署 \| 根據 \| 海水中 \| 大腸桿菌 \| 水平 \| 將 \| 泳灘 \| (學習) \| 水質 \| 評級 |
| | Gloss | environmental-protection-department \| according-to \| in-sea-wather \| E.coli \| level \| put \| beach \| (aptitude) \| water-quality \| grade |
| | English | The Environmental Protection Department grades the beaches according to the level of E.coli in the water |
| 32 | Chinese | 小班教學 \| 有助 \| 解決 \| 學習 \| (泳灘) \| 差異 |
| | Gloss | small-classes \| helpful \| solve \| aptitude \| (beach) \| difference |
| | English | Small classes may help solve the problem of aptitude differences |
| 33 | Chinese | 香港 \| 唯一的 \| 資源 \| 是 \| 人才 \| (意外) |
| | Gloss | hong-kong \| only \| resource \| is \| human-resources \| (accident) |
| | English | The only resource in Hong Kong is human resources |
| 34 | Chinese | 無論 \| 發生 \| 多少 \| 意外 \| (人才) \| 都是 \| 大家 \| 所 \| 不願見 \| 的 |
| | Gloss | irrespective \| happen \| how-many \| accident \| (human-resources) \| all-be \| we \| of \| not-wish-to-see \| AS |
| | English | Irrespective of the number of accidents, no one will wish to see any accidents happening |
| 35 | Chinese | 內地 \| 製作的 \| 電視 \| 劇集 \| (議員) \| 在 \| 本港 \| 甚受 \| 歡迎 |
| | Gloss | mainland \| of-production \| television \| drama-series \| (councillor) \| in \| hong-kong \| very \| popular |

Continuation of Table C.1

| | | |
|---|---|---|
| | English | Mainland-produced television drama series have become very popular in Hong Kong |
| 36 | Chinese | 選民 \| 將於 \| 區議會 \| 選舉中 \| 選出 \| 議員 \| (劇集) \| 加入 \| 區議會 |
| | Gloss | voter \| will-in \| district-councils \| election \| elect \| councillor \| (drama-series) \| join \| district-councils |
| | English | Voters will elect for members for each of the constituencies in the District Councils |
| 37 | Chinese | 他的 \| 英勇 \| 行為 \| 促使 \| 警方 \| 拘捕 \| 疑匪 \| (貢獻) |
| | Gloss | his \| bravery \| behaviour \| help \| police \| round-up \| suspect \| (contribution) |
| | English | His extraordinary bravery has helped the police round up the suspects |
| 38 | Chinese | 香港的 \| 經濟 \| 有賴 \| 中小型 \| 企業的 \| 貢獻 \| (疑匪) |
| | Gloss | of-hong-kong \| economy \| rely-on \| small-medium-sized \| enterprise \| contribution \| (suspect) |
| | English | The economy of Hong Kong is driven by small to medium sized enterprises |
| 39 | Chinese | 貧窮的 \| 村民 \| 有望 \| 入住 \| 設有 \| 水電的 \| 新居 \| (秩序) |
| | Gloss | poor \| villager \| hopefully \| live-in \| equipped-with \| water-electricity \| new-home \| (order) |
| | English | The poor villagers at last were given the hope of living in new homes equipped with electricity and water supplies |
| 40 | Chinese | 市民 \| 欣賞 \| 煙花 \| 時 \| 應 \| 遵守 \| 秩序 \| (新居) |
| | Gloss | citizen \| watch \| fireworks \| during \| should \| obey \| order \| (new-home) |
| | English | Members of the public are reminded to keep order during the fireworks display |
| 41 | Chinese | 由於 \| 進行 \| 水務 \| 維修 \| 部分 \| 樓宇的 \| 鹹水 \| (建議) \| 供應 \| 將 \| 暫停 |
| | Gloss | due-to \| conduct \| water \| alteration \| part-of \| of-building \| sea-water \| (suggestion) \| supply \| will-be \| suspend |
| | English | Sea water supply to some premises will be suspended for carrying out alteration work to water mains |
| 42 | Chinese | 只有 \| 那些 \| 具備 \| 理據的 \| 建議 \| (鹹水) \| 才會 \| 獲得 \| 支持 |
| | Gloss | only \| those \| fully \| justified \| suggestion \| (sea-water) \| will-be \| gain \| support |
| | English | Only cases that are considered fully justified will be supported |
| 43 | Chinese | 博物館 \| 展出 \| 多位 \| 國畫 \| 大師的 \| 作品 \| (問題) |
| | Gloss | heritage-museum \| exhibit \| many \| ink-painting \| of-master \| pieces \| (problem) |
| | English | Important works by ink painting masters are on show at Heritage Museum |
| 44 | Chinese | 我們 \| 現正 \| 面對 \| 經濟 \| 轉型 \| 所 \| 產生的 \| 就業 \| 問題 \| (作品) |
| | Gloss | we \| now \| facing \| economy \| restructure \| of \| caused \| employment \| problem \| (pieces) |
| | English | What we are facing is the employment problem caused by economy restructuring |

Continuation of Table C.1

| | | |
|---|---|---|
| 45 | Chinese | 議員 \| 就 \| 如何 \| 推動 \| 資訊 \| 科技 \| 發展 \| 向 \| 政府 \| (數量) \| 提供 \| 意見 |
| | Gloss | councillor \| about \| how \| drive \| information \| technology \| development \| to \| government \| (amount) \| provide \| advice |
| | English | Councillors advise the government on how to drive information technology development in Hong Kong |
| 46 | Chinese | 過去 \| 五年 \| 跨境 \| 旅客 \| 數量 \| (政府) \| 增長 \| 超逾 \| 一倍 |
| | Gloss | past \| five-years \| cross-boundary \| passenger \| amount \| (government) \| growth \| over \| doubled |
| | English | In the past five years, cross boundary passenger flow has more than doubled |
| 47 | Chinese | 科技的 \| 發展 \| 為 \| 大家 \| 帶來 \| 方便 \| (三歲) |
| | Gloss | of-technology \| development \| for \| we \| bring-about \| convenience \| (three-year-old) |
| | English | The development in information technology has brought about great convenience to people |
| 48 | Chinese | 接受 \| 學前 \| 教育的 \| 最低 \| 年齡 \| 應 \| 維持於 \| 三歲 \| (方便) |
| | Gloss | receive \| pre-primary \| of-education \| minimum \| age \| should \| remain-at \| three-year-old \| (convenience) |
| | English | The minimum age for pre-primary education should remain at 3 |
| 49 | Chinese | 纜車 \| 在 \| 有 \| 需要時 \| 將 \| 增加 \| 班次 \| 以 \| 應付 \| 乘客 \| (議案) \| 需求 |
| | Gloss | tram \| during \| have \| necessity \| will \| increase \| frequency \| to \| meet \| passenger \| (legalization) \| demand |
| | English | Tram services will be increased to meet passenger demand if necessary |
| 50 | Chinese | 議員 \| 將 \| 辯論 \| 一項 \| 議案 \| (乘客) |
| | Gloss | councillor \| will \| debate \| a-motion \| legalization \| (passenger) |
| | English | Councillors will debate a motion on the legalization |
| 51 | Chinese | 舉辦 \| 籌款 \| 活動的 \| 團體 \| 應 \| 採取 \| 預防 \| 措施 \| (海外) \| 防止 \| 舞弊 |
| | Gloss | organise \| fund-raising \| of-campaign \| organisation \| should \| take \| preventive \| policy \| (overseas) \| prevent \| fraud |
| | English | Organisers of fund-raising campaigns should take precautions against any possible fraud |
| 52 | Chinese | 警方 \| 相信 \| 該批 \| 毒品 \| 準備 \| 在 \| 本地 \| 及 \| 海外 \| (措施) \| 市場 \| 分銷 |
| | Gloss | police \| believe \| that-batch \| drugs \| ready-for \| in \| local \| and \| overseas \| (policy) \| market \| distribute |
| | English | Police believed the illicit drugs would be for the distribution in both local and overseas market |
| 53 | Chinese | 隊員 \| 專業 \| 的 \| 表現 \| (貨船) \| 對 \| 行動的 \| 成功 \| 有 \| 決定性 \| 作用 |
| | Gloss | team-member \| professional \| AS \| performance \| (cargo-vessel) \| to \| of-operation \| success \| have \| vital \| function |

Continuation of Table C.1

| | | |
|---|---|---|
| | English | Team members' professionalism played a vital role in the success of the operation |
| 54 | Chinese | 水警 \| 將 \| 該艘 \| 貨船 \| (表現) \| 截停 |
| | Gloss | marine-police \| put \| that \| cargo-vessel \| (performance) \| intercept |
| | English | Marine police intercepted the cargo vessel |
| 55 | Chinese | 勞工處 \| 密切 \| 跟進 \| 有關 \| 個案 \| 確保 \| 工傷 \| 僱員 \| (傳統) \| 獲得 \| 補償 |
| | Gloss | labour-department \| closely \| assist \| about \| case \| ensure \| industrial-injury \| employer \| (tradition) \| receive \| compensation |
| | English | The Labour Department is assisting in the case of industrial injury to ensure the injured parties receive the entitlements |
| 56 | Chinese | 供養 \| 父母 \| 是 \| 優良 \| 的 \| 傳統 \| (僱員) |
| | Gloss | support \| parents \| is \| honoured \| AS \| tradition \| (employer) |
| | English | It is a time-honoured tradition for children to support their parents when they get old |
| 57 | Chinese | 隨著 \| 內地 \| 市場 \| 開放 \| 香港 \| 將有 \| 很大 \| 的 \| 優勢 \| (滅火) \| 和 \| 發展 \| 空間 |
| | Gloss | with \| mainland \| markey \| expansion \| hong-kong \| will-have \| great \| AS \| advantage \| (fire-fighting) \| and \| development \| space |
| | English | With the expansion of the mainland market Hong Kong will gain great advantages and opportunities for development |
| 58 | Chinese | 早期 \| 警隊 \| 除 \| 負責 \| 維持 \| 治安 \| 外 \| 更 \| 兼負 \| 滅火 \| (優勢) \| 工作 |
| | Gloss | early-days \| police \| not-only \| responsible-for \| enforce \| law-and-order \| other-than \| also \| also-responsible-for \| fire-fighting \| (advantage) \| work |
| | English | In the early days, the police was responsible for enforcing law and order as well as fighting fire |

Table C.2: Handedness questionnaire

## Handedness Questionnaire

Participant Name: _____

Experimenter Name: _____

The motor functions listed below are usually completed by a single hand. Please tell us which hand do you use for completing these tasks. Please circle your answers in this 5-point-scale table.

| **Tasks** | **Left Hand Always** | | **Half-and-Half** | | **Right Hand Always** |
|---|---|---|---|---|---|
| Writing | 1 | 2 | 3 | 4 | 5 |
| Drawing | 1 | 2 | 3 | 4 | 5 |
| Throwing a ball | 1 | 2 | 3 | 4 | 5 |
| Holding "chopsticks" | 1 | 2 | 3 | 4 | 5 |
| Holding a hammer | 1 | 2 | 3 | 4 | 5 |
| Brushing teeth | 1 | 2 | 3 | 4 | 5 |
| Holding a pair of scissors | 1 | 2 | 3 | 4 | 5 |
| Opening a door | 1 | 2 | 3 | 4 | 5 |
| Striking a match | 1 | 2 | 3 | 4 | 5 |

Table C.3: Instruction sheet

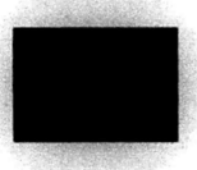
## Participant Instruction Sheet

On behalf of The Language Engineering Laboratory I thank you very much for participating in our experiment.

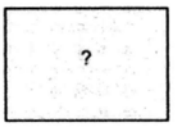
In this experiment, you are going to ***read silently*** sentences (默讀一些句子) that will appear on the LCD screen. Your task is to judge whether the sentences are acceptable to you (評估句子是否合理) or not ***according to your own impression***. There is no right or wrong answer.

The sentences will be presented as a ***sequence of slides***. At the beginning of a trial please fix your eye gaze on the "+" sign. This fixation screen will look like this:

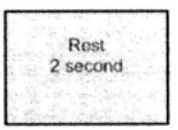
The first word of a sentence will appear soon after the disappearance of the "+" sign: Subsequent words will follow.



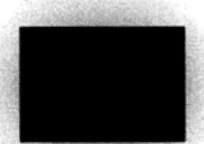
After the end of a sentence, as indicated by another fixation screen (with a "+" sign), you will be prompted to ***make a judgement*** on whether the sentence you just read is acceptable to you or not:




You will response by pressing a button on a computer mouse.
After that, you will then be given a 2-second break before the next sentence appears.




You are advised to ***maintain your eye gaze*** on the centre of the screen and ***try not to move your head or your body*** whenever the screen is with a black background. Like this:



There are altogether 10 sessions, each will last for about 3 minutes. You will be given enough time to rest and rélax between sessions.

# References

Allison, T. (1984). Recording and interpreting event-related potentials. In E. Donchin (Ed.), *Cognitive Psychophysiology: Event-related potentials and the study of cognition*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: Time course and scalp distribution. *Journal of Cognitive Neuroscience*, *11*(3), 235–260.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Borovsky, A. & Elman, J. L. (2006). Language input and semantic categories: A relation between cognition and early word learning. *Journal of Child Language*, *33*, 759–790.

Chatrian, G. E., Lettich, E., & Nelson, P. L. (1985). Ten percent electrode system for topographic studies of spontaneous and evoked EEG activity. *American Journal of EEG Technology*, *25*(2), 83–92.

Christiansen, M. & Chater, N. (1999a). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205.

Christiansen, M. H. & Chater, N. (1999b). Connectionist natural language processing: The state of the art. *Cognitive Science*, *23*(4), 417–437.

Christiansen, M. H., Conway, C. M., & Curtin, S. (2005). Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. In J. W. Minett & W. S.-Y. Wang (Eds.), *Language acquisition, change and emergence: Essays in evolutionary linguistics* (pp. 205–249). Hong Kong: City University of Hong Kong Press.

Christiansen, M. H. & Devlin, J. T. (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *Proceedings of the 19th Annual Cognitive Science Society Conference*, (pp. 113–118)., Mahwah, NJ. Lawrence Erlbaum.

Connolly, J. F. & Philips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, *6*(3), 256–266.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Elman, J. L. (1999). The emergence of language: A conspiracy theory. In B. MacWhinney (Ed.), *The Emergence of Language* (pp. 1–27). Hillsdale, NJ: Lawrence Earlbaum Associates.

Elman, J. L. (2001). Connectionism and language acquisition. In M. Tomasello & E. Bates (Eds.), *Language development: The essential readings* (pp. 295–306). Malden, MA: Blackwell Publishing.

Elman, J. L. (2003). Generalization from sparse input. In *Proceedings of the 38th Annual Meeting of the Chicago Linguistic Society*.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301–306.

Federmeier, K., Mai, H., & Kutas, M. (2005). Both sides get the point: Hemispheric sensitivities to sentential sonstraint. *Memory and Cognition*, *33*(5), 871–886.

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. In S. Pinker & J. Mehler (Eds.), *Connections and symbols* (pp. 3–71). Cambridge, MA: MIT Press.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the First Mind Articulation Project Symposium* (pp. 95–126). Cambridge, Mass.: MIT Press.

Gibson, E. & Pearlmutter, N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences*, *2*(7), 262–268.

Grüning, A. (2006). Stack-like and queue-like dynamics in recurrent neural networks. *Connection Science*, *18*(1), 23–42.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441.

Hald, H. A., Bastiaansen, M. C. M., & Hagoort, P. (2006). EEG theta and gamma responses to semantic violations in online sentence processing. *Brain and Language*, *96*(1), 90–105.

Hauk, O., Ford, M. H., Pulvermüller, F., & Marslen-Wilson, W. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage*, *30*(4), 1383–1400.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*, 301–307.

Hauk, O. & Pulvermüller, F. (2004). Neurophysiological distinction of action words in the fronto-central cortex. *Human Brain Mapping*, *21*(3), 191–201.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). New Jersey: Prentice Hall.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Hinojosa, J. A., Martín-Loeches, M., Muñoz, F., Casado, P., & Pozo, M. A. (2004). Electrophysiological evidence of automatic early semantic processing. *Brain and Language*, *88*(39-46).

Hsiao, F. & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*, 3–27.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *11th European Association for Lexicography International Congress*.

Kutas, M. & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470.

Kutas, M. & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M. & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Landi, N. & Perfetti, C. A. (2007). An electrophysiological investigation of semantic and phonological processing in skilled and less-skilled comprehenders. *Brain and Language*, *102*(1), 30–45.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the n400. *Nature Reviews Neuroscience*, *9*(12), 920–933.

Lee, C.-Y., Tsai, J.-L., Chiu, Y.-C., Tzeng, O. J.-L., & Hung, D. L. (2006). The early extraction of sublexical phonology in reading Chinese pseudocharacters: An event-related potentials study. *Language and Linguistics*, *7*(3), 619–636.

Liu, Y., Perfetti, C. A., & Hart, L. (2003). ERP evidence for the time course of graphic, phonological, and semantic information in Chinese meaning and pronunciation decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1231–1247.

Luck, S. J. (2005). *An Introduction to the event-related potenntial technique*. Cambridge, MA: The MIT Press.

Ma, X. (2004). Hong Kong Parallel Text. *Linguistic Data Consortium, Philadelphia, LDC2004T08*.

Marcus, G. F. (1998). Can connectionism save constructivism? *Cognition*, *66*, 153–182.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Marslen-Wilson, W. & Tyler, L. K. (1975). Processing structure of sentence perception. *Nature*, *257*(5529), 784–786.

Martín-Loeches, M., Hinojosa, J. A., Casado, P., Muñoz, F., & Fernández-Frías, C. (2004). Electrophysiological evidence of an early effect of sentence context in reading. *Biological Psychology*, *65*, 265–280.

McClelland, J. L. & Plaut, D. C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences*, *3*(5), 166–168.

McClelland, J. L. & Rumelhart, D. E. (1986). *Parallel distributed processing: Psychological and biological models*, volume 2 of *Computational models of cognition and perception*. Cambridge, MA: MIT Press.

Meng, X., Tian, X., Jian, J., & Zhou, X. (2007). Orthographic and phonological processing in Chinese dyslexic children: An ERP study on sentence reading. *Brain Research*, *1179*, 119–130.

Minsky, M. L. & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, Mass.: MIT Press.

Penolazzi, B., Hauk, O., & Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology*, *74*, 374–388.

Perfetti, C. A., Liu, Y., Fiez, J., Nelson, J., Bolger, D. J., & Tan, L.-H. (2007). Reading in two writing systems: Accommodation and assimilation of the brain's reading network. *Bilingualism: Language and Cognition*, *10*(2), 131–146.

Perfetti, C. A. & Tan, L. H. (1998). The time course of graphic, phonological, and semantic activation in Chinese character identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *241*(1), 101–118.

Pizzagalli, D. A. (2007). Electroencephalography and high-density electrophysiological source localization. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Phychophysiology* (3rd ed.). New York: Cambridge University Press.

Proverbio, A. M., Vecchi, L., & Zani, A. (2004). From orthography to phonetics: ERP measures of grapheme-to-phoneme conversion mechanisms in reading. *Journal of Cognitive Neuroscience*, *16*(2), 301–317.

Redington, M. & Chater, N. (1998a). Connectionist and statistical approaches to language acquisition: A distributional perspective. In K. Plunkett (Ed.), *Language acquisition and connectionism* (pp. 129–191). Hove, UK: Psychology Press.

Redington, M. & Chater, N. (1998b). *Connectionist and statistical approaches to language acquisition: A distributional perspective*, volume 13, (pp. 129–191).

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469.

Rodriguez, P., Wiles, J., & Elman, J. L. (1999). A recurrent neural network that learns to count. *Connection Science*, *11*(1), 5–40.

Rodriguez, R. (2001). Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation*, *13*(9), 2093–2118.

Rohde, D. L. T. & Plaut, D. C. (1999). Simple recurrent networks can distinguish non-occurring from ungrammatical sentences given appropriate task structure: Reply to marcus. *Cognition*, *73*(3), 297–300.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations. In D. E. Rumelhart, J. L. McClelland, & University of California San Diego PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, volume 1 (pp. 319–362). Cambridge, MA: MIT Press.

Rumelhart, D. E. & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, volume 1 of *Computational models of cognition and perception*. Cambridge, MA: MIT Press.

Ruz, M. & Nobre, A. C. (2008). Attention modulates initial stages of visual word processing. *Journal of Cognitive Neuroscience*, *20*(9), 1727–1736.

Snyder, P. J. & Harris, L. J. (1993). Handedness, sex, and familial sinistrality effects on spatial tasks. *Cortex*, *29*(1), 115–134.

Tan, L. H., Spinks, J. A., Gao, J.-H., Liu, H.-L., Perfetti, C. A., Xiong, J., Stofer, K. A., Pu, Y., Liu, Y., & Fox, P. T. (2000). Brain activation in the processing of Chinese characters and words: A functional MRI study. *Human Brain Mapping*, *10*(1), 16–27.

Tomasello, M. (2001). The item-based nature of children's early syntactic development. In M. Tomasello & E. Bates (Eds.), *Language development: The essential readings* (pp. 169–186). Malden, MA: Blackwell Publishing.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tzeng, O. J.-L., Hung, D. L., & Wang, W. S.-Y. (1977). Speech recoding in reading Chinese characters. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(6), 621–630.

Tzeng, O. J.-L. & Wang, W. S.-Y. (1983). The first of two R's. *American Scientist*, *71*, 238–243.

Valian, V., Prasada, S., & Scarpa, J. (2006). Direct object predictability: Effects on young children's imitation of sentences. *Journal of Child Language*, *33*(02), 247–269.

van den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience*, *13*(7), 967–985.

van der Velde, F. (2005). Modelling language development and evolution with the benefit of hindsight. *Connection Science*, *17*(3-4), 361–379.

van der Velde, F. & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*, 37–108.

van der Velde, F., van der Voort van der Kleij, G. T., & de Kamps, M. (2004). Lack of combinatorial productivity in language processing with simple recurrent networks. *Connection Science*, *16*(1), 21–46.

Wang, W. S.-Y. (1973). The Chinese language. *Scientific American*, *228*, 50–60.

Wong, F. C. K., Minett, J. W., & Wang, W. S.-Y. (2006). Reassessing combinatorial productivity exhibited by simple recurrent networks in language acquisition. In *2006 International Joint Conference on Neural Networks (IJCNN)*, (pp. 2905–2912)., Vancouver, Canada.

Wong, F. C. K. & Wang, W. S.-Y. Finding early brain signature to contextual influences during sentence comprehension: An ERP study on Chinese reading.

Wong, F. C. K. & Wang, W. S.-Y. (2007a). Combinatorial productivity through the emergence of categories in connectionist networks. *Dynamics of Continuous, Discrete and Impulsive Systems (Series A): Advances in Neural Networks*, *14*(SI), 650–657.

Wong, F. C. K. & Wang, W. S.-Y. (2007b). Generalisation towards combinatorial productivity in language acquisition by simple recurrent networks. In *2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems. KIMAS '07: Modeling, Exploration, and Engineering*, (pp. 139–144)., Waltham, MA, USA.