

The role of social-pragmatic cues
in word learning: a neural
network model

Gregory A. Caza

a thesis submitted for the degree of
Master of Science
at the University of Otago, Dunedin,
New Zealand.

30 June 2008

Abstract

Human infants begin to produce speech at the beginning of the second year of life. Some theories propose that language is acquired by simply recognising statistical properties in linguistic input. Other hypotheses consider the interactive environments in which humans are raised, looking for links between emerging social skills and word learning.

The social-pragmatic theory of language acquisition suggests that the foundation of word learning is an ability to read the intentions of another, especially the intention to communicate (Akhtar and Tomasello, 2000). The theory posits a special role for the ability to respond to joint attention, a form of intention recognition which has been shown to facilitate word learning (Baldwin, 1993). Social-pragmatic theory argues that the cognitive abilities that develop in the second year of life are not merely coincidental, but are an essential component of language acquisition.

This first goal of this thesis is to develop a neural network model to investigate word learning by implementing key concepts of the social-pragmatic theory. Cognitive skills that develop in the second year of life are shown to facilitate word learning, with the model reproducing characteristics of the 'vocabulary spurt' that can be seen around 18 months of age (O'Grady, 2005). The second goal is to relate the model to current neurobiological research. The neural correlates of intention recognition and lexical retrieval are tentatively defined, permitting a discussion of the brain regions common to both processes. The prefrontal cortex, in particular, is discussed to investigate how its general functions could be harnessed by mechanisms for word learning and intention recognition. One novel contribution of the model is to tie together joint attention and word learning using a reward-based learning scheme.

Acknowledgements

First and foremost, I must thank my supervisor, Dr. Alistair Knott. Not only did he help all of my ideas to make sense, but he kept all my sea lions in check.

I would also like to thank Edwin van der Ham, who gave me a wee bit of code to bootstrap my stimulus-response network. Thanks also go out to the other lecturers and students of the Artificial Intelligence research group. They challenged my hypotheses, focused my thoughts, talked to me about highly-trained monkeys, or just distracted me with a game of darts when I really needed it.

I dedicate this thesis to my father, Bryan Caza. He may not have known a lot about neurolinguistics, but he was a master at sharing my attention.

Contents

1	Introduction	1
2	Literature Review: Experimental Studies	4
2.1	First Language Acquisition	4
2.1.1	Statistical Learning	5
2.1.2	Social-Pragmatic Theory	7
2.2	The First Year of Life	8
2.2.1	Intention Recognition in the First Year of Life	10
2.2.2	Joint Attention in the First Year of Life	11
2.3	The Second Year of Life	12
2.3.1	Intention Recognition in the Second Year of Life	13
2.3.2	Joint Attention in the Second Year of Life	14
2.4	Intention Recognition and Language	14
2.5	Discussion	18
3	Literature Review: Computational Models	20
3.1	Statistical Learning Models	21
3.1.1	A Model of Gaussian Distribution	21
3.1.2	A Cross-Situational Inference Model	22
3.1.3	The Lexicon as Exemplars Model	24
3.1.4	The Simple Recurrent Network Model	26
3.2	Models of Social-Pragmatic Theory	29
3.2.1	A Gaze Following Model	29
3.2.2	Yu and Ballard’s Unified Model	32
3.3	Discussion	34
4	Joint Attention as a Filter on Word Learning	35
4.1	Background	35
4.2	Experiment A: The Simple Word Learning Model	37
4.2.1	Introduction	37
4.2.2	Methods	37
4.2.3	Results	39
4.2.4	Discussion	40
4.3	Experiment B: Word Learning with a Joint Attention Filter	40
4.3.1	Introduction	40
4.3.2	Methods	44

4.3.3	Results	46
4.3.4	Discussion	46
4.4	Experiment C: Development of Joint Attention	48
4.4.1	Introduction	48
4.4.2	Methods	48
4.4.3	Results	49
4.4.4	Discussion	52
4.5	General Discussion	55
5	A Neurobiologically-Inspired Model of Intention Recognition and Word Learning	56
5.1	Neural Substrates of Noun and Verb Concepts	56
5.2	Neural Substrates of Intention Recognition	60
5.2.1	Background: a model of PFC function	60
5.2.2	Intention Representation in the PFC	63
5.2.3	The Mirror System: action recognition	65
5.2.4	Intention Recognition and Neurolinguistics	68
5.3	A Schematic Model of Action Execution and Recognition	71
5.4	An Intention Recognition Model of Word Learning	74
5.5	Experiment D: Learning Nouns and Verbs	77
5.5.1	Methods	78
5.5.2	Results	79
5.5.3	Discussion	79
5.6	Experiment E: Learning When to Enable the Joint Attention Filter . .	80
5.6.1	Introduction	80
5.6.2	Methods	80
5.6.3	Results	83
5.6.4	Discussion	84
5.7	Experiment F: Bootstrapping the Joint Attention Filter	84
5.7.1	Introduction	84
5.7.2	Methods	85
5.7.3	Results	86
5.7.4	Discussion	87
6	Conclusion	89
6.1	Evaluation	89
6.2	Future Work	92
	References	94
A	A Model of Plan Competition	105
A.1	Background	106
A.1.1	Serial Behaviour	106
A.1.2	Associative Memory	107
A.1.3	Backward Inhibition	108
A.2	Experiment G: The Plan Competition Network	110

A.2.1	Design	111
A.2.2	Training Parameters	111
A.2.3	Preliminary Results	112
A.2.4	Discussion	113

List of Tables

3.1	Summary of results from the unified model (Yu and Ballard, 2007). . .	33
4.1	The top row represents visual, sensorimotor information. The bottom row shows examples of coinciding emphasised speech. The single congruent case is highlighted.	36
4.2	SWLM training results for different problem sizes, showing the number of epochs required to map the vocabulary with 100% accuracy.	39
4.3	Illustration of the joint attention model. The top two rows display visual information. The bottom row shows the contiguous word held in the phonological loop. The crucial trigger is MOTHER-TALK, and the subsequent learning opportunity is highlighted.	41
4.4	Estimating the percentage of communicative events involving joint attention (based on Harris, Jones, Brookes, and Grant, 1986).	43
5.1	Sample training input (first column) and the mappings that the noun SWLM would try to learn (second column.) Essentially, the network will be instructed that DOG maps to <i>dog</i> more often than it sees it mapping to any particular verb.	78
5.2	A sample timeline for learning when to enable the JAF.	81
5.3	Network parameters used in Experiment E.	82
5.4	Network parameters used in Experiment F.	86

List of Figures

2.1	Sample illustration of word learning with a ‘vocabulary spurt’ at 18 months. (Adapted from O’Grady, 2005.)	13
3.1	A simple model of lexical acquisition (from Siskind, 1996).	23
3.2	The LEX model (from Regier, 2005).	25
3.3	An example SRN (from Elman, 1990). The dotted connections are trainable, whereas the recurrent ones are fixed at unit weight.	27
3.4	A visual overview of the gaze following model (adapted from Triesch, Teuscher, Deák, and Carlson, 2006.)	30
3.5	Comparison of association probabilities from the unified model. (Adapted from Yu and Ballard, 2007.)	33
4.1	A reduced representation of the Simple Word Learning Model, implemented as a backpropagation neural network with one hidden layer. . .	38
4.2	An extension to the SWLM. The original neural network (Figure 4.1) is gated by a preceding network that implements the joint attention filter. Thicker lines between the layers indicate stronger connections.	45
4.3	Experiment B. A comparison of learning speed with (<i>JA</i>) and without (<i>SL</i>) joint attention filtering.	46
4.4	Mean number of epochs required to learn the 20 words in Experiment B, with and without the JAF. Error bars represent one standard deviation. . .	47
4.5	Experiment C1. Results with CDS equal to 0.10.	50
4.6	Experiment C2. Results with CDS equal to 0.20	50
4.7	Experiment C3. Results with CDS equal to 0.60.	51
4.8	Experiment C4. Results with CDS equal to 0.80.	51
4.9	Mean number of epochs required to learn the vocabulary in Experiment C, for different values of CDS. Error bars represent one standard deviation. . .	52
4.10	Mean number of epochs in Experiment C, for the smaller values of CDS, to learn all twenty words (<i>Full</i>) or the five before the filter is activated (<i>Pre-filter</i>).	53
4.11	Examples of learning with CDS 0.10, from vocabulary size 1 to 10. (<i>i</i>) Statistical learning only; (<i>ii</i>) learning with joint attention.	54

5.1	A lateral view of the major lobes of the human brain: frontal (yellow), parietal (green), occipital (blue), and temporal (red) (adapted from Gathercole, 1999). Numbers indicate Brodmann areas, with the prefrontal cortex (PFC) comprised of 8, 9, 10, 11, 44, 45, 46, and 47. Regions of interest in the PFC include dorsolateral (9 and 46), orbitofrontal (10 and 11), and Broca's area (44 and 45). Other relevant locations include the premotor cortex (6), medial temporal gyrus (21), inferior temporal gyrus (20), and Wernicke's area (posterior 22).	57
5.2	A diagram of simple stimulus-response behaviour (based on Miller and Cohen, 2001). Each stimulus (S1, S2, S3) can be followed by one or more voluntary responses (R1 and R2). Thicker lines indicate stronger connections. Intermediate, hidden units represent processing pathways between the two.	61
5.3	An illustration of the guided activation theory (based on Miller and Cohen, 2001). PFC units (P1, P2) excite hidden units to bias the selection of a desired response.	62
5.4	A simple model of an integrated action execution (AE) and action recognition (AR) system. Each node in the PMC represents a possible action.	67
5.5	An example of action execution: the intention to jump.	68
5.6	An example of action recognition: recognising that a conspecific intends to jump.	68
5.7	An adaptation of Miller and Cohen's GAT, combined with the AE/AR system. The PFC, IT, and PMC layers are connected by hidden layer. Inhibitory connections between hidden neurons are not shown.	71
5.8	Example behaviour of the system for a representation of SELF-JUMP.	73
5.9	Example behaviour of the system for recognising that a DOG is performing the action JUMP.	75
5.10	Illustration of an intention recognition word learning network with a joint attention filter and two simple word learning modules.	76
5.11	Learning nouns and verbs in Experiment D, presented as number of words learned over successive epochs.	79
5.12	Experiment E. Learning when to filter, based on reinforcement from pre-trained word learning networks. A training example is considered 'correct' when it produces the expected behaviour in the network.	83
5.13	Mean results for Experiment F.	87
A.1	A competitive queuing model that demonstrates contention scheduling (based on Houghton and Hartley, 1995). Arrow-headed connections are excitatory; other connections are inhibitory.	106
A.2	A Hopfield Net (Hopfield, 1982). All connections are shown as excitatory for the purposes of illustration, but may become inhibitory during training.	108
A.3	The expected behaviour of a Hopfield Net. A: Two patterns are presented during training; dashed connections are weakened, thick connections are strengthened, and other connections are left unmodified. B: After training, (i) a subset of one of the learned patterns is presented and (ii) the network settles into an activation that completes it.	109

- A.4 A simplified representation of the Plan Competition Network. Black connections are of fixed weight, while grey connections vary in strength. The dotted line represents a gate on the inhibitory connections. . . . 110

Chapter 1

Introduction

Language, quite simply, is a window through which we can reach out and touch each other's minds (Altmann, 1997, p. 233).

How do children first learn language? Decades of research have failed to provide a definitive answer to this puzzling question. Many theories have been proposed, but they often fall short of explaining every aspect of the process. It is important to remember that the skills involved in developing language are themselves experiencing developmental change as the child ages (Bloom, 2000), suggesting that children of different ages may use different strategies for word learning. Also, existing knowledge of language informs future attempts at acquisition (Baddeley, Gathercole, and Papagno, 1998). Once word learning begins, there are likely a number of interacting skills and factors that come into play to facilitate the process (Woodward, 2000).

One important question in the field of language acquisition concerns the role of social-pragmatic cues when children are learning their very first words. **Social-pragmatic cues** are the non-verbal signals that accompany spoken language in human communication, including eye gaze and pointing gestures. Just as these signals help adults to communicate with foreign-language speakers, they can help infants to understand that there is an intention to communicate *before* they can fully comprehend the words being spoken. Infants must learn that verbal communication *exists* before they can begin to learn how use language. A great deal of research has studied the influence of these social cues, specifically how recognising them may facilitate word learning (e.g., Baldwin, 1993; Sabbagh and Baldwin, 2005; Tomasello, 2003). The aim of this thesis is to use artificial neural network models to simulate how recognising communicative intentions is a benefit to the process of single word acquisition. In order to do this, I will propose a relationship between the cognitive skills that children develop in the

second year of life—especially between 18 and 24 months—and the changes that can be seen in their word learning strategies.

The key idea in my model is that an infant’s mechanism for word learning can be ‘switched on’ or ‘switched off’. I will implement the process that enables or disables this mechanism in the form of a filter that is influenced by social-pragmatic cues, and measure its effect on the acquisition of single word meanings. My proposal is that the infant eventually learns to enable the word learning mechanism when a communicative intention is recognised. I will also model this learning process of identifying *when* to turn on the filter.

This thesis begins with an interdisciplinary literature review. In Chapter Two, I review relevant experimental research. Studies from developmental psychology and linguistics allow me to elaborate on my hypothesis. The results of those studies also provide guidance for the form and parameters of the computational models I will describe later. In the third chapter, I focus on the computational models of other researchers. Models that take social cues into account are contrasted with those that do not. Considering the advantages and disadvantages of these models will generate a list of criteria for assessing my own simulations.

The fourth chapter is used to tie together the research that I will have discussed up to that point. I begin to support my hypothesis by presenting several experiments that model word acquisition using neural networks. Word acquisition in this context is considered to be the process of learning the correct mapping between a sensorimotor concept and a lexical term. A sequence of models is given, beginning with one based on a simple and well-documented architecture. I then progressively add the features necessary for my model to highlight the role of social cues in early word learning. The main extension in Chapter Four will be to implement a filter that can enable or disable the word learning mechanism.

By controlling the word learning, a state which can either be active or inactive, the filter is used to ignore irrelevant ambient speech and other situations comprised of noise. In short, the filter enables word learning at appropriate times and inhibits it other times. I experiment with the effect of turning this filter on at different times in the timeline of vocabulary acquisition. My simulations show some of the behaviour of real language learners, including a ‘vocabulary spurt’ (O’Grady, 2005).

In Chapter Five, I take a step back and begin to look at my model of word acquisition from the perspective of cognitive neuroscience. Imaging and lesion studies are reviewed to make tentative proposals about the neural substrates of the cognitive pro-

cesses which I am modelling. The review involves a discussion of different brain regions, as well as a high-level model of action and intention recognition. These proposals will allow me to make further incremental extensions to my model. In the extended model, I consider how the word learning filter *learns* when to turn itself on. I propose that this learning is via a reward-based reinforcement scheme, which I believe is a novel implementation in computational models of vocabulary acquisition.

In the last chapter, I will draw conclusions based on my research and suggest steps for future work.

In Appendix A, I present a side project that could take my model one step further. I will include details of one more experiment, beginning to explore these possibilities.

Chapter 2

Literature Review: Experimental Studies

This chapter will provide an overview of the experimental literature on language learning. I will begin by summarising two main theories of language acquisition. I will then assess these theories from the perspective of what is known about development in the first year of life (which is largely prelinguistic) and in the second year (when language abilities develop rapidly.) As I do so, I will elaborate on why I find one particular theory, known as Social-Pragmatic Theory, to be the more convincing explanation of how and why language is learned. I will then discuss possible connections between language and other cognitive abilities that emerge during the same period of development.

2.1 First Language Acquisition

The mechanism of first language acquisition is a complicated topic and has perplexed researchers for decades. Even after a massive quantity of research, there are many unanswered questions. Why are humans the only species to develop language? Why does language learning appear so effortless for infants? Why does production seem to consistently begin at a specific point in development—around 12 months of age—and often follow a predictable progression? This final question is the most important one for the current work. I am interested not so much in the root origins of language as I am in the role that cognitive development plays in facilitating and shaping the process of acquisition. In order to analyse language acquisition within a developmental framework, I will first contrast two theories: statistical learning and the social-pragmatic theory.

2.1.1 Statistical Learning

Statistical learning is one main category of modern theories of language acquisition; it borrows from the earlier theory of behaviourism by highlighting the importance of association learning. Behaviourists (like B. F. Skinner) suggested that children learn language through simple associative learning and then make generalisations via a process of induction (see Tomasello, 2003, pp. 2–3 for a more detailed review). One of the main criticisms of the behaviourist view is that some elements of grammar are too abstract to be learned in such a manner. At present, there are very few researchers who would support behaviourism as a complete explanation of how language is learned. However, some elements of the theory survive as components of modern alternatives like statistical learning.

According to statistical learning theory, first language acquisition is essentially a numbers game. Infants learn the semantics and syntax of language by using the experience gained from a large input set of examples. It might be child-directed speech, or just what is overheard, but an infant has access to a plethora of spoken utterances in the course of an average day. According to this theory, established skills of pattern recognition and statistical inference allow for conclusions to be reached about mappings between words and objects in the environment. The general supposition is that language comprehension begins when the infant has acquired enough experience (i.e., observed enough examples) to make the correct associations.

At the individual word-learning level, the theory proposes that associations can be made by hypothesising about the set of potential meanings and finding the consistent element of the non-linguistic contexts in which that word is heard (Yu and Ballard, 2007). Siskind (1996) referred to this process as ‘cross-situational inference.’ For example, pretend that I do not know the meaning of the word “truck”, but that I encounter it used in a number of situations: I see a child playing with a toy vehicle and she uses the word “truck”; I see a program on television about a “Monster Truck Rally”; I am about to cross the street and someone tells me to “watch out for that truck”; and so on. According to statistical learning theory, I am able to extract the common element from these scenes—a powerful, four-wheeled vehicle—and associate it with the uttered word “truck”. This argument seems very reasonable and makes intuitive sense, but can it account for every aspect of language acquisition?

Of course, statistical learning theory is not focused merely on drawing inference from context. Part of the appeal of statistical learning theory is that it does not depend on special, language-dedicated skills or regions of the brain. Many find the notion that a

dedicated language system evolved suddenly for humans, with no apparent precursors in non-human primates, rather improbable (see also Ghazanfar, 2008). According to statistical learning theory, the kinds of inferences that are proposed as a method for acquiring language are just one application of domain non-specific learning abilities. The theory posits that infants (and older language learners) deploy cross-situational cognitive skills of memory, attention, and pattern recognition.

Saffran, Aslin, and Newport (1996) found evidence that 8-month-olds are capable of segmenting nonsense words from an auditory stream. The segmentation of words is certainly a very necessary skill for comprehending language. The infants were habituated to a continuous-speech stream of four three-syllable nonsense words in random order; transitional probabilities between syllables were higher within words than between words, to give some clue as to what constituted a word boundary. During the test phase, the infants were presented with ‘words’ drawn from the training vocabulary and ‘non-words’ made up of the same syllables. There was a significant dishabituation to the non-words versus the words, suggesting that the infants distinguished between them and found the non-words to be novel stimuli. It is unlikely that this task was performed using linguistic knowledge for two reasons. First, the subjects were prelinguistic. Second, the words were artificial, rather than drawn from any natural language to which the infant might have been exposed.

Saffran *et al.*’s data are very compelling, but are such statistical abilities *sufficient* for language learning? My pet dog can learn to associate the utterance “walk” with a fun time outside, to the point that I can embed it in an irrelevant sentence and she will still get excited when she hears it. The dog is clearly segmenting my speech, but it is highly unlikely that she has grasped the semantics of the word “walk”. Evidence suggests that non-human primates, such as tamarins (as reviewed by Saffran, 2003), can also distinguish word boundaries while demonstrating no ability to actually use language or produce speech.

Statistical learning theory is supported by many computational models (e.g., Elman, 1990; Regier, 2005; McMurray, 2007) that show the ability of a computer to process language-like stimuli without any sort of innate module or domain-specialised skills. However, just because a computer can solve the problem space by applying statistical principles does not mean that is how an infant actually accomplishes the same task. Useful analogies aside, a computer is *not* a human brain. More importantly, computers are not reared in a socially interactive environment. Statistical learning theory does not factor in the environment in which children are raised, leading to what Akhtar and

Tomasello (2000) refer to as a “neglect of the social dimensions of word learning.”

2.1.2 Social-Pragmatic Theory

The **social-pragmatic theory** of language acquisition proposes that an understanding of communicative situations forms the foundation of word learning (Akhtar and Tomasello, 2000). This comprehension of the intentions behind the utterances of others is one form of **intention recognition**. It has been conceptualised under a few different names, including ‘intention-reading’ (Tomasello, 2003) and ‘mind reading’ (Baron-Cohen, 1995). Intention recognition skills include: sharing attention with others, following the attention of others, actively directing others’ attention, and recognising the intentions underlying observed actions (Tomasello, 2003, p. 3). Understanding the communicative intention of another requires an ability to identify and interpret his or her **referential intent**, whether it takes an explicit form (e.g., pointing) or an implicit one (e.g., eye gaze.) Interpreting this intent involves identifying the referent of his or her communication. For example, if a mother wants to direct her son’s attention to a dog, she might say “Look at the doggie” and point in the direction of the dog. In this case, the referent would be the dog and her son would have to follow the direction of her gesture to find the dog that she is talking about.

According to Tomasello (2003, p. 4), intention recognition skills “are very likely unique to human beings, and they probably emerged relatively recent in human evolution.” There is evidence that chimpanzees can follow gaze direction and perhaps understand a little about intention in *action* (Tomasello, Call, and Hare, 2003). Other species that have shown an ability to learn to follow human eye gaze and gesturing include domestic dogs (Hare, Brown, Williamson, and Tomasello, 2002) and dolphins (Marino, Connor, Fordyce, Herman, *et al.*, 2007). Even domestic goats appear to be capable of following human pointing gestures and the gaze of conspecifics—although perhaps not human eye gaze (Kaminski, Riedel, Call, and Tomasello, 2005). However, humans are the only species that seem to master the *complete* set of skills—especially the directing of attention and the cultural learning of intentional actions. If this is true, it is only sensible to look for possible connections between intention recognition skills and other uniquely-human attributes, like language.

Social-pragmatism proposes that it is not enough to bombard an infant with overheard speech until they reach some critical mass of experience, as suggested by statistical learning theories. Indeed, the environment in which a child is raised is inherently noisy, presenting him or her with many conflicting and incorrect examples of language

use. A purely statistical learning strategy would face an arduous task to find the necessary patterns in the spoken words heard by the child.

Social-pragmatic theory does not dismiss the associative abilities of infants. In fact, Tomasello suggests that associative and pattern-finding abilities are demonstrated so early in life that children should be able to produce language much earlier in development if all they had to do was to form associations. Rather, he argues that there are much *better* learning mechanisms that children employ to learn language, but that these skills do not develop until around the first birthday. To Tomasello (2003, p. 90), words are learned almost as a “by-product of social interaction”; language develops when it does because that is the earliest the infant has the ability to grasp the necessary concepts. Skills of intention recognition are essential to forming this kind of understanding. Thus, there is a dependence on the cognitive skills that are necessary to appreciate what language is trying to convey and—more importantly for my research—identify when a social-communicative situation exists that will be conducive to word learning. To explain why the social-pragmatic theory is compelling, I will now turn to a review of literature that supports it. First, I will summarise studies that find evidence in the first year of life of precursors for communication and the relevant social skills. Following that, I will report on experiments demonstrating the influence of the cognitive skills that emerge during the second year of life.

2.2 The First Year of Life

Children babble from a young age (Tomasello, 2003, p. 35) and use non-verbal forms of communication, such as pointing, from about nine months of age (Baron-Cohen, 1989). However, as mentioned earlier, they do not utter their first recognisable words until around 12 months of age; this milestone has great cross-cultural consistency (Tomasello, 2003, p. 19). The **critical period** hypothesis suggests that neural plasticity for language learning declines from its peak at 1 year of age until early adolescence, and then even more sharply throughout adulthood (Newport, 1990). However, many researchers argue that there is no clearly delineated critical period, suggesting that maturation brings about a continuous decline in language acquisition ability (Tomasello, 2003, pp.286–287).

How the very first word is learned by an infant, with no prior knowledge of language is a difficult question, commonly referred to as a **bootstrapping** problem. Of course, it is inappropriate to reach conclusions about language comprehension by assuming it

begins with the first spoken word. Production skills seem to lag behind other language abilities in development, even once the infant begins to speak (Tomasello, 2003, pp. 79–80). For instance, 12-month-olds appear to comprehend 5–10 times as many words as they are able to produce themselves (as cited in Woodward, 2000).

Prelinguistic infants display many language-related skills. Burnham and Dodd (2004) investigated the ‘McGurk effect’, in which a speech syllable dubbed over an incongruent lip movement will be auditorially-perceived as a combination of the two when they are presented together (i.e., lip movements for [ga] dubbed with [ba] will be perceived as [da] or [ɖa].) In adults, the effect is seen as evidence for multi-modal speech processing that takes into account visual and auditory information (McGurk and MacDonald, 1976). In Burnham and Dodd, 4½-month-olds were habituated to McGurk-dubbed (experimental group) or congruent (control group) auditory/visual stimuli. During the test phase, all infants were presented with different auditory stimuli corresponding to the three possible percepts. Infants in the experimental group showed a significant preference for the fused syllables ([da] or [ɖa]) over the one that was actually heard ([ba]), suggesting that they were affected by the visual information; control infants showed no significant preference. Four- and five-month-olds also prefer to look at faces when the apparently-articulated vowel is congruent with an auditorially-perceived one, versus incongruent articulating faces (as reviewed by Liberman and Mattlingly, 1985). These results support the hypothesis that even prelinguistic infants will use visual information when it is available and integrate it with auditory information in a manner similar to adults.

Many other experiments have explored the auditory abilities of prelinguistic infants. Tomasello (2003, p. 30) reviewed studies indicating the ability to find abstract auditory patterns at 7 months, and that even newborns can associate visual experiences with separate auditory ones. Intriguingly, other primates are also capable of finding patterns in speech-like stimuli (Tomasello and Akhtar, 2003). Therefore, there is evidence that prelinguistic infants and non-human primates can find patterns in auditory information, a necessary skill for the comprehension of spoken language. Why then, are these two populations not capable of producing language? The simplest answer is that extracting abstract patterns is a necessary skill for language, but is certainly not a *sufficient* one.

Prelinguistic infants also appear to have a special affinity for listening to speech. Those as young as 9 months recognise it as a special kind of sound (see Woodward, 2000 for a review). Finally, studies have shown that 12-month-olds are able to extract emotional information from intonation and rhythm of speech before they are able to

comprehend its meaning (see Baldwin and Moses, 1996 for a review of several studies).

It is important to note that many of the prelinguistic studies just mentioned can also be—and have been—viewed as evidence for the statistical learning theory. Abstract auditory pattern recognition and visual association are examples of abilities that would logically be important to language acquisition, regardless of the mechanism. It is necessary to view the language learning process in the larger context of cognitive development to illuminate the appeal of the social-pragmatic theory. Despite the evidence for an earlier awareness of language, speech still does not begin until around the second year of life. What development, then, does an infant lack at 5 or 7 or even 9 months of age that generally prevents word production?

2.2.1 Intention Recognition in the First Year of Life

As mentioned above, recognising a speaker’s communicative intentions is key to the social-pragmatic theory of language acquisition. 9- to 12-month-olds display some skills of intention recognition by sharing attention, following attention and gestures, directing attention using gestures, and imitating the intentional actions of others (see Tomasello, 2003, pp. 21–28 for a review). As Tomasello so elegantly explained it, infants begin “to use adults as social reference points.” It is not considered a coincidence that these crucial skills are entrenched around the first birthday. But, when do these skills begin to develop, and what are their precursors?

Woodward (1998) studied whether 6- to 9-month-olds are sensitive to the goals of an action. Subjects were presented with two toys placed on individual pedestals. The experiment consisted of a habituation trial where an arm would enter from outside the field of view and grasp one of the toys; a test trial followed, with the grasp of a different toy on the same side, or the same toy on a different side. When a human arm performed the grasping, the infants would dishabituate to a new-toy/same-side trial (i.e., new goal), but not to a same-toy/new-side one. In Woodward’s estimation, the change of side in movement was more perceptually salient because the agent was standing closer to one of the toys; but, the change in toy was what caused the dishabituation, suggesting the infants encoded their observations based on some aspect of the situation beyond mere saliency. This finding suggests that the infants classified the grasp as a goal-directed action, dishabituating when they observed a new goal for the action in the same location but not when the *same* goal was witnessed in a different location. Interestingly, the infants showed no significant preference when the experiment was performed with a mechanical grasper rather than a human arm. Thus, even as young

as 5 months of age, humans appear to distinguish between intentional human action and simple physical motion.

2.2.2 Joint Attention in the First Year of Life

The second necessary skill, as identified by Tomasello, is the ability to establish what he refers to as ‘joint attentional frames’ between caregiver and child. A joint attentional frame is often defined (Eilan, 2005) as a state in which two individuals are attending to the same object *and* both are aware of the fact that they are attending to the same object. Thus, two people independently looking at the same object is a necessary—but not sufficient—condition for the establishment of these joint attentional frames. According to this strict definition, the joint attentional frame is a triadic interaction with an implicit mutual awareness that adds an interactive dimension. The triadic dimension assumed by Tomasello and Eilan is not always considered necessary by other researchers. To reduce confusion, I will adopt the term **joint attention** to describe the broad cases where one’s attention is intentionally directed to be congruent with another, and **shared attention** for the subset of cases that meet the additional constraint of mutual awareness (see also Triesch *et al.*, 2006).

Obviously, visual joint attention (hereafter referred to simply as ‘joint attention’) is not possible unless the infant can perceive the caregiver (e.g., mother)¹ and successfully judge the target of her attention. Beyond the ability to simply notice an agent, joint attention requires a reading of *what* the other is attending to. At the early stages of language acquisition, an infant will obviously not be able to use verbal cues to discern the caregiver’s target. Thus, there is a dependence on other referential cues such as pointing gestures and/or eye gaze.

Very basic eye-based gaze following can occur as early as 3 months of age (Hood, Willen, and Driver, 1998). Electroencephalography (EEG) studies of event-related brain potentials showed that children as young as four months differentially process the gaze of an adult, depending on whether or not it is object-directed (Hoehl, Reid, Mooney, and Striano, 2007). There are limits, however, on the quality of gaze following in young infants. Butterworth and Jarrett (1991) found that 6-month-olds would look in the correct direction but not necessarily to the exact location, tending to fixate on the first object scanned; 12-month-olds more correctly attended to the correct object, but only 18-month-olds would tend to follow gaze to objects behind them, outside their

¹From now on, for the sake of illustrating examples, I will assume that the caregiver is female and that the infant is male. No gender bias is intended.

initial visual field. There are other forms of referential intent, such as pointing and other hand gestures. Again, 10- to 12-month-olds will look to the correct target of pointing when it is nearby, but only 15- to 18-month-olds can do so for objects farther away (as reviewed by Thoermer and Sodian, 2001).

Woodward (2003) studied 7- to 12-month-olds following an adult's gaze to one of two toys. A similar paradigm was used to the grasping experiment mentioned previously. The difference here was that they viewed an adult looking at one of two toys, rather than watching a disembodied arm performing a grasping action. Although all the infants tended to follow the adult's gaze, it was only the 12-month-olds that dishabituated when the adult looked at a new toy on the same side. Such a result is significant, because it means that the older children did not dishabituate for just any change in the scene; a change in toy was more interesting than a change in location. The toy, itself, was not the salient item because it was present throughout the trial. Rather, it was the deployment of gaze to a new object—but not a new location—that was considered interesting. This distinction suggests that a connection was drawn between the adult and the specific object she was looking at. The gaze study stands in contrast to the grasping study cited above, in which 6-month-olds seemed to draw a connection between the action and the goal. Thus, it is only between 9 and 12 months that infants seem to form a representation of the relationship between the gazer and the target of his/her gaze.² Before this understanding is formed, gaze following must be motivated by another goal, such as seeking novel stimuli (e.g., Triesch *et al.*, 2006), or based on some innate orienting response (see Woodward, 2003 for a review.)

Again, it is difficult to accept that such skills would operate in total isolation of other interactive and observational processes like language acquisition. Indeed, there is evidence that early affinity for joint attention is correlated with early language comprehension and production (Carpenter, Nagell, and Tomasello, 1998). Because correlation does not necessarily imply causation, however, it is important to also consider how language skills also improve in later stages of cognitive development.

²The results of Thoermer and Sodian (2001) would suggest that the representations formed by 12-month-olds may not be as robust as indicated by Woodward's data. However, the former study did show clear developmental progress between 10 and 12 months of age, suggesting that an understanding of such relationships is formed at or around 12 months of age.

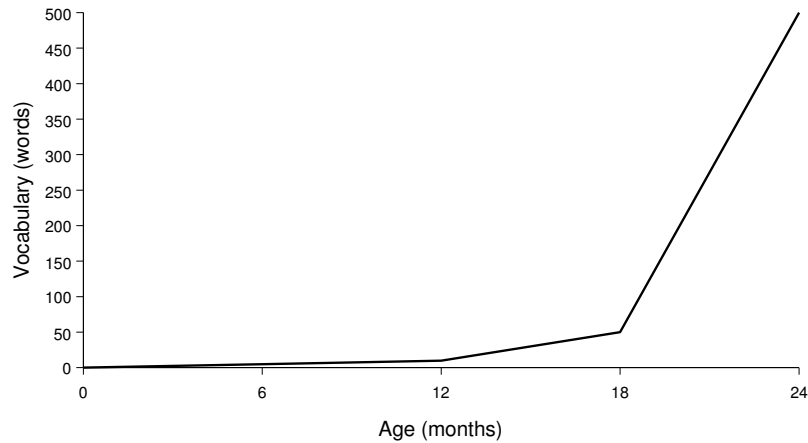


Figure 2.1: Sample illustration of word learning with a ‘vocabulary spurt’ at 18 months. (Adapted from O’Grady, 2005.)

2.3 The Second Year of Life

As a child approaches 24 months of age, words can sometimes be learned in a single exposure (Regier, 2005). Many researchers believe this skill, known as fast or **rapid mapping**, is one of the markers of a ‘naming explosion’, ‘vocabulary burst’, or **vocabulary spurt**. According to the vocabulary spurt hypothesis, the rate of word learning shows a sharp increase around 18 months of age, or approximately when a vocabulary of 50 words is reached (see Figure 2.1). However, other researchers dispute the existence of a vocabulary spurt—or at least the prevalence of one. They argue for a steady, more gradual increase with no clear point of inflection in word learning rate (see also Tomasello, 2003, p. 50). Ganger and Brent (2004) performed a longitudinal study of the rate of vocabulary acquisition by children in their second year of life. They plotted rate of word learning against time, and then attempted to fit the data with a logistic function (i.e., indicating a spurt) or with a quadratic one (indicating a gradual increase.) Their analysis suggests that possibly only 18% of children demonstrate a measurable spurt. However, one criticism (see Li, Zhao, and MacWhinney, 2007) of their study was their focus on 12- to 20-month-olds might not have covered a large enough period to allow all of the children to demonstrate a spurt.

Whatever the case, researchers generally agree that word learning starts slow and picks up speed throughout the second year of life. This same developmental span also

marks a number of cognitive milestones. Is this a total coincidence, a mere correlation, or do certain cognitive skills predict the acceleration?

2.3.1 Intention Recognition in the Second Year of Life

Intention recognition abilities continue to develop throughout the second year of life. Meltzoff (1995) studied imitation in 18-month-olds to assess their intention reading abilities. Infants in a control group witnessed a novel object manipulation, repeated three times. For an ‘intention’ group, adults were observed making three unsuccessful attempts at completing the action, providing no verbal or facial feedback to indicate failure. Upon imitation, both groups of infants successfully completed the *intended* act 80% of the time, with little or no trial-and-error. However, a similar experiment with a machine demonstrating the actions did not see children reproducing the target acts with any reliability. Thus, there is evidence that 18-month-olds are able to recognise intended acts, and also that they differentiate between human intentions and mechanical motion.

2.3.2 Joint Attention in the Second Year of Life

Between 12 and 24 months, infants also improve at other judgements of referential intent. By 18 months, a child can reliably follow gaze based on eye movements alone (see Triesch *et al.*, 2006 for a review). It has been suggested that, in the second year of life, there is a shift from focusing on perceptual cues, like object salience, to a reliance on more social cues, like eye gaze (Bloom, 2000).

Mundy and Newell (2007) drew a sharp distinction between responding to joint attention and initiating joint attention, two abilities which develop in parallel. The latter is defined as the infant’s use of his own referential intent to direct the attention of another, and it also develops throughout the second year of life (Bates, Thal, Whitesell, *et al.*, 1989). Such intent to spontaneously share experience is itself a form of communication. As mentioned above, this ability—also known as ‘(proto)declarative pointing’—is considered to be unique to humans (Tomasello, 2003). Moreover, ability to initiate and respond to joint attention at 12 and 18 months has been shown to be predictive of language ability at 24 months (Mundy, Block, Delgado, Pomares, *et al.*, 2007).

2.4 Intention Recognition and Language

A link between intention recognition and language is seen by many as crucial because “language is nothing more than another type—albeit a very special type—of joint attentional skill; people use language to influence and manipulate one another’s attention” (Tomasello, 2003, p. 21).

That children correctly learn words in a joint attention situation is remarkable. The mother cannot simply explain what she is trying to achieve. When explaining a concept, we often define words in terms of other words. But, such a strategy is not possible when the listener has a small or non-existent vocabulary. If the mother says something to the effect of “Look, a toma”, it is a trivial matter to turn to and see what she is looking at, inferring that whatever novel object is present is probably called a toma—with enough experience, that is. These steps seem superficially simple to an adult, but of course they will not be so obvious to an infant. The child must invoke a number of cognitive processes, including: recognising that the noises coming out of her mouth are not just random, inferring some sort of communicative intent from her speech, realising that the direction of her gaze is relevant to what she is saying, and assessing the direction and target of her gaze. Only then can he form a representation of what he is seeing and hearing to create the appropriate links and learn the intended information. It is good news indeed that children do not name objects simply by association because up to half of what their parents say to them refers to things that are currently outside of their attention (Harris *et al.*, 1986).

To understand how children overcome these obstacles, it is important to consider how they learn to connect speech with referential intent throughout the second year of life. 12- to 18-month-olds will orient toward speakers for many sorts of vocalizations, like sighs, but have proven more likely to actually follow a speaker’s gaze after a referential utterance (summarised in Sabbagh and Baldwin, 2005). Children past 15 months of age will tend to look towards the sounds of a speaker’s voice in a labelling situation, even if the speaker is not visible (Baldwin, Markman, Bill, Desjardins, *et al.*, 1996). This action appears to be more than reflexive, however; 12-month-olds are more likely to follow speaker gaze when there are two possible referents rather than a single ambiguous one (Sabbagh and Baldwin, 2005). And, while younger infants may judge any sound as an object label, 20-month-olds are discriminating enough to only do so with speech (summarised in Woodward, 2000).

Oft-cited work by Baldwin (1991) provides insight into how infants in the second

year of life learn to use non-verbal cues to identify the referents of ambiguous speech. 16- to 19-month-olds were studied using different conditions in which a novel label was assigned to one of two novel toys. In what I will refer to as the *correlated* condition, the experimenter uttered the label while she and the infant were looking at the same toy. The correlated case represents a joint, or even *shared*, attention situation because the experimenter would glance at the infant to ensure that he or she was focused on the toy in question. For the *uncorrelated* condition, the experimenter and the infant were looking at different toys at the time of the utterance. In the uncorrelated case, joint attention was not possible because the experimenter focused on an object inside a bucket and avoided looking directly at the infant. In the test phase, the infants were asked to find the toy corresponding to the novel label. It was considered a ‘mapping error’ if children in the uncorrelated group produced the toy that had been in their focus when the label was uttered, rather than the one on which the experimenter had been focused. When tested, infants from the correlated group chose the correct toy significantly more than would be expected by chance; this result suggests that they correctly assigned the label to the toy. Overall performance for infants in the uncorrelated group was at chance, demonstrating that they avoided possible mapping errors by not assigning the label to either toy. When considered separately, the *older* children (18- to 19-month-olds) in the uncorrelated group did show a better-than-chance ability to also choose the correct toy—i.e., the one that had been in the bucket. Although joint attention was not possible in that case, the experimenter’s gaze provided a clear, non-verbal cue regarding her referential intent. A number of alternative explanations for the results, such as toy preference, level of engagement, comprehension of familiar labels, etc. were controlled for in the experimental design or ruled out during data analysis. Clearly, the infants did not simply associate the word they heard with whatever they were looking at when they heard it. Rather, the data suggest that they used the non-verbal cue of gaze to help them identify the correct referent.

Part of the appeal of Baldwin’s experiment lies in its simplicity. Later research has continued to strengthen the argument for the role of social cues in word learning by investigating possible alternative explanations, as well as teasing out some of the possible confounds in the original study. Baldwin (1993) was a similar study, with 19- to 20-month-olds, in which both toys were placed in opaque containers. The experimenter looked into one of the containers and uttered the novel label and then the toys were removed from the containers, sequentially. The toy the experimenter looked at was

either removed first (correlated), or second (uncorrelated). Temporal contiguity was controlled for by imposing a delay of 10 seconds between the extractions. A second experiment controlled for saliency by having the experimenter perform a *non*-referential act—playing with the lid while looking at the infant—intended to enhance the saliency of one of the containers. In every case, the infants were significantly more likely to assign the label to the toy indicated by the experimenter’s gaze, suggesting that they gave referential cues priority over competing factors like temporal contiguity or object salience.³ In a third study, Baldwin *et al.* (1996) situated one of the speakers behind a screen, making it impossible for the baby to follow his gaze or other visual cues. 18- to 20-month-olds did not assign object labels in that situation, even though a second, visible adult would simultaneously engage in joint attention on the object. Apparently, a situation of joint attention was not sufficient by itself. The joint attention was not with the actual speaker and, because the speaker could not be seen, the referential intent was not clear.

Of course, judgement of referential intent is not the only weapon in the arsenal of a word learner. Many different skills can be brought to bear in different situations. Samuelson and Smith (1998) devised a scheme where 18- to 24-month-olds played with three different distractor objects and then moved across the room to play with the target object on a glittery, cloth-topped table. The target object was then mixed with the distractors and a novel object label was uttered; the children were more likely to assign the label to the target object. The point of the experiment was to show that “general attentional and memorial processes” aided word learning, rather than knowledge about referential intent. It is difficult to conclude that this was proven because it is debatable whether or not there was clear referential intent in the process—at the very least, any intent was not in *conflict* with object saliency. The experiment does suggest that, in some circumstances, saliency is evidently sufficient to promote word learning.

Woodward and Hoyne (1999) used a joint attention paradigm to contrast 13-month-olds with 20-month-olds. During training, joint attention was established and a novel object was either correlated with a spoken word or with a novel sound (e.g., an electronic beeper.) When tested, both groups of children had learned the spoken word as the object label. (It is assumed the younger children were also able to do so because

³Moore, Angelopoulos, and Bennett (1999) suggested this skill may not be reliable until closer to 24 months; but, more recently, Houston-Price, Plunkett, and Duffy (2006) demonstrated that 15-month-olds may show a preference for gaze over at least saliency of movement.

there was no distractor object, as in Baldwin (1991).) However, only the 13-month-olds, and not the 20-month-olds, adopted the sound-object mapping. While it is uncertain whether the younger children were classifying the novel sound as a label, per se, a couple of conclusions do seem clear. First, the experiment gives evidence that associative abilities have developed by the second year of life. Second, older children—who are presumably as good or even better at association than the younger ones—have learned to employ other strategies and will even give them priority over simple association in communicative situations.

Akhtar has performed experiments examining language learning from overheard words. In one experiment (Akhtar, Jipson, and Callanan, 2001), 2-year-olds either played with a novel toy while interacting with an experimenter or they observed an assistant playing with the toy. In both cases, the experimenter assigned the toy a verbal label during play. Children proved equally able to learn labels in both conditions. In a later experiment (Akhtar, 2005), another condition was tested in which the child was presented with a distractor toy before the overhearing situation was staged. Again, subjects ranging in age from 2 to 3 years appeared equally able to learn a novel object label in the different conditions.

At first glance, these results would seem contrary to some of the tenets of the social-pragmatic theory of word learning. Certainly, it is difficult to argue that the strict, triadic definition of shared attention was satisfied; overhearing a conversation does not require that the speaker and the listener have a *mutual awareness* of attending to the same object. However, conditions were sufficient for the less stringent case of joint attention and I would argue that the spirit of the social-communicative act is very much intact. The infant is not randomly looking in the same direction as the adult. He is looking at the novel object because the adult is providing clues, like gestures and gaze, that the object is worth attending to. Whether or not the speaker has an awareness that the observer is actually directing his attention according to her indications does not appear to be relevant, at least in this case. From the listener's perspective, the referential intent is quite clear: "she wants me to look at that thing over there, while she's saying 'toma'." Otherwise, the label could easily have been assigned to the distractor toy, to the assistant, or just not learned at all. It is possible that the shared attention described by Tomasello (2003) is necessary for younger children, but perhaps joint attention is sufficient by the end of the second year of life.

To summarise, there is evidence of infants assigning special attention to speech and recognising non-verbal cues as carrying relevant information. That, in short, "18- to

24-month-olds filter their word learning through their understanding of human action and, more specifically, communication” (Woodward, 2000). Is it a coincidence that language production begins in the second year of life, when these other important social-cognitive skills also develop? One hypothesis suggests that initial language acquisition may actually help encourage the development of intentional understanding (Sabbagh and Baldwin, 2005). As Baldwin and Moses (1996) suggest, “a growing understanding of language could well be a catalyst for the acquisition of new notions of others as sources of information.”

2.5 Discussion

Association and statistical learning theories do not seem sufficient to account for the complicated process of language acquisition. I have reviewed excellent evidence that suggests word learning is aided by the ability to comprehend and interpret non-verbal cues, as asserted by the social-pragmatic theory. These cues include eye gaze and gesturing, as well as the communicative intentions of another.

Accepting that word acquisition is facilitated by the ability to recognise communicative intention still leaves many open questions. How do the abilities interact? What is the developmental trajectory? What are the brain regions involved? I will try to answer some of these questions by implementing and experimenting with neural network simulations that exploit social cues in word learning situations. In the next chapter, I will lay the foundation for my design by reviewing the computational models of other researchers.

Chapter 3

Literature Review: Computational Models

Before embarking on the development of a completely new computational model, it is important to consider previous work in the field. This chapter will summarise the models of other researchers in the areas of language learning and social-pragmatic cues. I will highlight the advantages and disadvantages of these models, with the goal of justifying why a new approach is warranted.

In order to provide a context for the discussion of these models, it is helpful to review some of the characteristics of real language learning that would ideally be emulated by an artificial simulation. These points, many of which were mentioned in the previous chapter, are summarised below.

1. *Bootstrapping*. Lexical acquisition must begin with little or no prior knowledge.
2. *Acceleration*. Learning begins slowly, but the rate of acquisition increases after some knowledge is accumulated. This acceleration can happen sharply in one step, or spurt, but does not necessarily do so for all language learners.
3. *Noise*. Word learning remains successful under circumstances that are much less than ideal. The system must be tolerant of incorrect and irrelevant input data.
4. *Rapid Mapping*. The meaning of a word can sometimes be learned in a single exposure.
5. *Development*. A child does not remain static throughout the learning process. Cognitive and physical development—including improvements to memory, perception, and understanding—happen in parallel with language acquisition.

The ways a child *uses* language also evolve over time (see also Tomasello, 2003, p. 185).

6. *Difficulty*. Not all words were created equal. The many factors that can increase the difficulty of learning a word include class, frequency of use, homonymity with other words, and ambiguity of meaning.
7. *Neurological Plausibility*. It has become increasingly important for computational neuroscience simulations to strive for a plausible model of the human brain.

3.1 Statistical Learning Models

3.1.1 A Model of Gaussian Distribution

McMurray (2007) aimed to demonstrate that any vocabulary spurt is not a function of “specialized processes” such as social cues, so much as an inevitable by-product of statistical acceleration. His model assumes that word learning occurs in parallel and that word learning difficulties (i.e., the amount of exposure needed to learn a word) can be approximated by a Gaussian distribution.

In McMurray’s model, words are assigned threshold difficulty values *a priori*. Each word is then given a point at every time-step of training, being considered learned when it accumulates enough points to pass its own threshold. The Gaussian Distribution model demonstrated a very characteristic spurt, with the rate of acquisition accelerating sharply after 2000 time-steps. Further simulations experimented with adding specialised mechanisms, wherein words already learned would either promote (i.e., add extra points to) or hinder (i.e., deduct points from) the learning of new words. The former was intended to model a form of fast-mapping, while the latter simulated possible interference from prior experience. In both cases, the slope of the rapid learning was attenuated but the spurt-like behaviour was still present.

The final simulations used real distributions of 2000 English words from adult-directed (ADS) and child-directed speech (CDS). In this case, a word’s difficulty was assumed to be inversely related to its frequency of utterance. The network was trained with either the ADS distribution or the CDS one and, in both cases, a vocabulary spurt was demonstrated.

Because of the consistent result of a vocabulary spurt across all simulations, McMurray argues that word learning is very parsimonious. It is proposed that some of the

characteristics of word learning are inevitable results of inherent statistical properties and that there is no need to search for further explanations of the nature or nurture variety.

The simplicity of the model is very appealing. It also appears to satisfy the constraints of bootstrapping and acceleration. One of its most reliable characteristics, however, is a vocabulary spurt. The occurrence of a spurt is *so* reliable, in fact, that it is never absent, even though experimental evidence shows us that not all real children demonstrate a spurt. While it could be argued that the simulation suitably models those children who *do* actually experience a spurt, its applicability to the general population (and possibly the majority—see Ganger and Brent, 2004) is reduced because of this behaviour.

Certain of the simplifications in the implementation make it difficult to draw too many conclusions from its behaviour. First of all, noise was never added to the idealised input data. Secondly, it is excellent that CDS is contrasted with ADS, but input from overheard (e.g., Akhtar *et al.*, 2001) or referentially-ambiguous (e.g., Baldwin *et al.*, 1996) speech was never considered.

Furthermore, the simulation is in no way dynamic. Beyond the natural mathematical properties of accumulated knowledge, word learning continues at a constant pace. It is important to acknowledge that such a fact is at the root of the paper’s argument and the author does allow for the possibility that other factors might help the learning process. However, it is problematic to conclude that other skills are absolutely unnecessary without assessing the plausibility of the input data. Finally, although words are given variable difficulty levels, those numbers are assigned based solely on rate of utterance and without regard to other contextual influences from the environment in which they are heard. It is interesting that all words are assumed to be learned in a linear and monotonically-improving fashion. Do the learners never get mixed up or make mistakes?

3.1.2 A Cross-Situational Inference Model

Siskind (1996) modelled cross-situational language learning techniques using a set of inference rules. As discussed in the previous chapter, cross-situational inference involves deducing the meaning of a word by gleaning the common elements from all observances of the word. Siskind formalises this goal as solving what he calls the **mapping problem**. The mapping in question is how the learner associates conceptual meaning with spoken utterances. How is the word ‘ball’ associated with the round red object, for

example, and not with the colour red or with all round shapes?

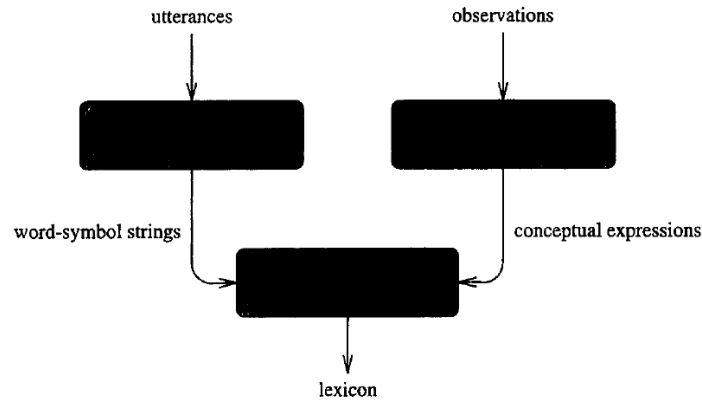


Figure 3.1: A simple model of lexical acquisition (from Siskind, 1996).

Siskind conceptualised word acquisition using the simple model shown in Figure 3.1. There are two streams of input: a speech perception one from spoken utterances and a perceptual/conceptual one from observations. There is a ‘speech-perception faculty’, which parses and segments the utterance stream into words. A second ‘faculty’ uses the perceptual/conceptual input to hypothesise about meanings for each of the lexical items in the other stream.

The way the model learns is by receiving an utterance, such as “John took the ball”, along with a set of hypothesised meanings for the different words in the utterance. The algorithm then uses its inference rules and any prior knowledge to map the individual words to meanings. In brief, the model performs its inference learning in two stages. Step one assembles a set of conceptual symbols representing a word symbol, using cross-situational information to reduce a ‘possible set’ of symbols until it converges to a ‘necessary set’ of symbols. In the second step, a similar process occurs that prunes a possible set of expressions of the symbols from step one until a single ‘conceptual expression’ remains.

The author tested with synthetic corpora of utterance-meaning pairs for the simulations, using a selection rule based on Zipf’s Law.¹ All input data were restricted to between 2- and 29-word utterances (mean 4.99 to 6.29), paired with meanings of 30 or less conceptual symbols. Beyond that, corpora could be tweaked according to size of vocabulary, rate of homonymy, and number of conceptual symbols.

¹If words are ranked in descending order of frequency, then the occurrence of a given word is inversely proportional to where it fits in this ranking (see Siskind, 1996); in other words, there is a very small number of words that appear frequently.

The model displayed characteristics that reflect actual word learning in children. Interestingly, some—but not all—of the simulations demonstrated a vocabulary spurt. Lexical acquisition was slow until approximately 25 words had been learned, at which point the rate accelerated. Words are converged on more quickly with more experience, as inferences can be made from prior knowledge. The model was also able to occasionally learn words in a single exposure, as seen with children starting around 24 months of age (Regier, 2005). The model even proved capable of performing reasonably well with levels of uniformly-distributed noise of up to 10%. There is little doubt that real-life word learning input is noisy, and Siskind rightly points out that it is difficult to quantify what would be a plausible amount. In Chapter 4, however, work I will review on mother-child interaction (Harris *et al.*, 1986) suggests that 10%—or even the 20% used in another experiment (Siskind, 1994)—might be a serious underestimation.

Siskind’s model is an improvement over McMurray’s in the sense that it successfully handles noise, it has a more layered (and possibly more plausible) implementation of word difficulty, and it does not universally demonstrate the acceleration principle. It also satisfies requirements for bootstrapping and rapid mapping. However, it still ignores any developmental influences on the system and makes no attempt to suggest neural correlates for its mechanisms. Finally, it does not predict any of the findings regarding the role of intention recognition in word learning.

3.1.3 The Lexicon as Exemplars Model

Regier (2005) created the Lexicon as Exemplars (LEX) computational model with the goal of addressing four specific developmental phenomena. The first was ease of learning, which is essentially the fast mapping mentioned above. The other three—phonological sensitivity (applying similar sounding words to different objects), shape bias (generalising a label to differently-shaped objects), and mutual exclusivity (learning two labels for the same object)—are beyond the scope of this thesis. The purpose of his model is to demonstrate that these developments can be explained using progressive associative and attentional skills, suggesting that conceptual insights are not necessary for improved word acquisition.

The LEX model is based on exemplars of form and meaning that cluster along relevant dimensions (e.g., voicing and shape, respectively.) It is a bidirectional multilayer network that aims to solve Siskind’s mapping problem from either side. If given a word form as input, its output is a probability distribution over possible meanings. If given a meaning as input, its output is a probability distribution over possible word forms.

The symmetrical architecture used consisted of two input/output layers and two hidden layers. As seen in Figure 3.2, there were associative weights to define the exemplar clusters and attentional weights that encoded dimensions of form and meaning; both sets were adjusted during training.

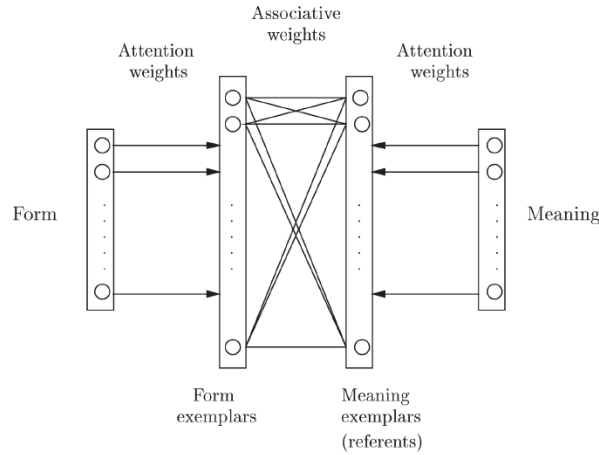


Figure 3.2: The LEX model (from Regier, 2005).

Over time, the network learns to allocate attention away from insignificant dimensions (e.g., pitch in English) to ones that are predictive of correct mappings (e.g., voicing of consonants in English.) This ability to identify language-specific features is an intriguing characteristic of the LEX model, especially considering theories of a critical period for language learning.

The LEX model satisfied its design constraints, with Regier suggesting that all four processes are mediated by the underlying principle of selective attention. The one process relevant to this thesis, fast mapping, was achieved admirably. Not only did the network demonstrate the ability, but it acquired it in a developmental fashion. Early on, it took 10 trials to learn a novel word and, over time, it became able to do so in just one trial. LEX also successfully models the principles of bootstrapping, acceleration, and difficulty. There is mention of a noise parameter in the mathematical definitions, but the author includes no discussion of any effects from it. Also, learning gets better over time, but it does so with no fundamental change to the network architecture. While such a lack of change in process is fundamental to the hypothesis of the LEX model, such an approach does not take into account the fact that an infant’s brain is actually changing throughout the second year of life.

Regier suggests that learning improvements modelled by LEX are the result of a “learner gradually determining which aspects of the world are relevant for commu-

nication and attending preferentially to those dimensions.” The notion that paying attention to what is important facilitates language use is superficially similar to social-pragmatic theory. However, what LEX models is not the sort of joint attention situations that are of interest to that theory. An insight into human intention is not necessary to realise that pitch is not a defining characteristic of the English language, for instance. Rather, Regier argues that the foundation for the developmental phenomena modelled by LEX is not a conceptual insight but rather selective attention. Like McMurray’s argument, this stance minimises any acknowledgement of an infant’s social environment.

Regier allows for the possibility that social cues such as intention understanding could be used to *guide* the selective attention simulated in the LEX model. He acknowledges that some non-human species are capable of associative learning (not to mention selective attention), even though they do not acquire language, and that “language *is* [emphasis his] inescapably referential.” Regier seems to agree with Tomasello that conceptual insights may indeed happen to jump-start initial word learning at 12 months, but he maintains that they are not important to fast mapping and other improvements in word learning through the second year of life. This argument would suggest there is a higher cognitive process in play at an early stage in development that is absent later in life, which is difficult to accept. If conceptual insight is a factor at twelve months of age, it is necessary to consider its influence at 18 and 24 months, as well.

3.1.4 The Simple Recurrent Network Model

Elman (1990) developed a seminal model to simulate language acquisition by a child. One of the goals was to suggest that developmental constraints may actually facilitate learning by buffering a network (child) from being overwhelmed by full, early exposure to a complex domain (language.)

Elman’s simulations employed different implementations of a **Simple Recurrent Network (SRN)**. In such a network, the hidden activation is fed back in the next time step to add context to subsequent inputs. An example is shown in Figure 3.3.

In the implementation, words are binary-encoded to form an input vector to the network. The output vector is decoded to produce a prediction of the next word in a sentence. Training was performed using three different schemata:

1. using all training data and a fully-developed network;
2. with the training data organized and presented with increasing complexity;

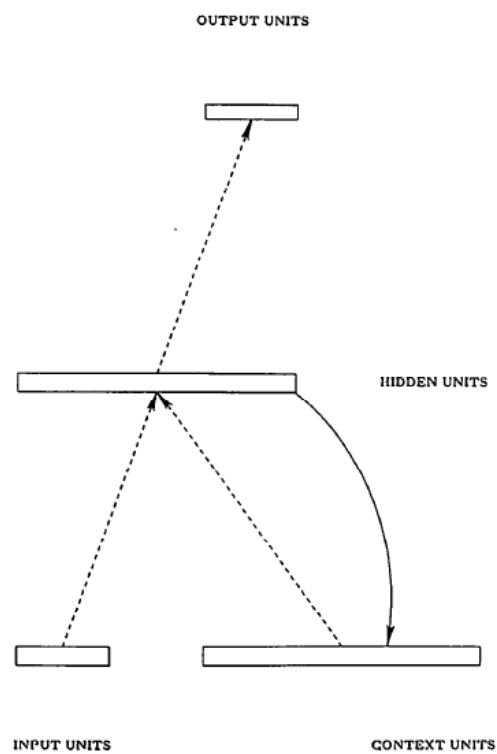


Figure 3.3: An example SRN (from Elman, 1990). The dotted connections are trainable, whereas the recurrent ones are fixed at unit weight.

3. beginning with a limited memory that increased throughout training.

When the network was initialised with full memory and exposed to the full data set, it could not master the task. Performance remained poor with the training data, without even considering its ability to generalise to novel input.

In the second case, Elman organized the training data by complexity. Each of the following distributions was presented, in order, for 5 epochs:

1. 10 000 simple sentences;
2. 7 500 simple and 2 500 complex sentences;
3. 5 000 simple and 5 000 complex sentences;
4. 2 500 simple and 7 500 complex sentences;
5. 10 000 complex sentences.

A low final error term indicated success, with the network also generalizing well to novel sentences.

In the third case, the network developed over the course of training, to simulate the growing memory and attention span of a child. The memory effect was simulated by resetting the recurrent feedback after a ‘memory window’ had elapsed. One of two lengths would be randomly selected and the context units would all be set to 0.5 after that word. The window lengths for the different training phases were as follows:

1. 3 or 4 words;
2. 4 or 5 words;
3. 5 or 6 words;
4. 6 or 7 words;
5. no memory window used.

The first phase of training was extended to 12 epochs to reach performance comparable to the other training regime. The other 4 phases used 5 epochs each. Error and ability to generalize at the end of training were as good as in the incremental input case.

Elman suggests that taking an evolving approach to the training is an effective way to simulate how a child is developing throughout the course of language acquisition.

The author hypothesizes that training with a subset constructs a “foundation for future success” by constraining the solution space such that there are less local error minima in which to trap the network. In effect, the simpler sentences (or lesser memory capacity) filter out “stimuli which may either be irrelevant or require prior learning to be interpreted.”

The SRN model meets the criterion of bootstrapping. It satisfies the difficulty requirement in the sense that neural networks inherently learn different patterns at different speeds, however the relative difficulties are essentially chosen at random and do not necessarily reflect reality. There is not enough detail in the paper to conclude that acceleration or rapid mapping behaviours were achieved in the simulations; the concept of noise is discussed, but not formally investigated. Most importantly, however, it is the first model explored here that considers developmental factors. The network evolves as it learns, with advancements that are not simply a by-product of accumulated knowledge. However, there is little effort on the author’s part to discuss the neural mechanisms of word learning in any detail.

3.2 Models of Social-Pragmatic Theory

In contrast to the many computational models of statistical learning, the equivalent literature on social-communicative influences is quite sparse. Before I describe one such model, I would like to describe a computational model of gaze following and joint attention. Although its context is not word learning situations, it models the developmental track of joint attention and its use of a reward-based learning scheme is intriguing.

3.2.1 A Gaze Following Model

Triesch *et al.* (2006) defined a ‘Basic Set’ that they view as sufficient for the emergence of gaze following: perceptual skills and preferences, reward-driven learning, habituation, and a structured social environment. They created a computational model of gaze following between an infant and a caregiver, defining infants as ‘pleasure-driven agents’ seeking interesting visual stimuli. All of the skills from their Basic Set are in place by 3-4 months of age and they are simulated to demonstrate progressive development of gaze following throughout the first year of life.

The author’s hedonistic classification of infant behaviour is important to their reward-driven learning proposal. The gaze-following mechanism is hypothesised to de-

velop as “infants gradually discover that monitoring their caregiver’s direction of gaze allows them to predict where interesting visual events will be.” The model is trained using a form of **reinforcement learning** known as **temporal difference learning** (TD learning). TD is a learning scheme with no explicit teacher, unlike supervised learning approaches. Rather, the system uses past experience to predict future behaviour (see also Haykin, 1999 for an elaboration.) TD is used to solve the ‘temporal credit assignment’ problem: if a sequence of actions brings favourable results, how can one know which actions in the sequence were more important to the outcome and assign credit appropriately? TD solves the problem by having a prediction of reward, with this prediction reevaluated over successive time steps. When a favourable outcome (or reward) occurs, a prediction signal is generated and propagated back through time to assign credit to the individual steps that caused it; an unfavourable result (or punishment) would likewise generate a signal to assign blame to previous time steps. ‘Good’ actions (i.e., those that bring about positive outcomes) are reinforced, making them more likely to occur in the future, whereas ‘bad’ actions are discouraged. Best of all, TD learning has some degree of neurological plausibility (Schultz, 1998).

Triesch *et al.*’s definition of habituation is also closely tied to their learning hypothesis. As an infant habituates to a stimulus and grows tired of looking it at, the process is classified as a declining reward value over time. Dishabituation to a change in stimuli is considered a recovery of reward. The TD algorithm, which aims to predict reward, consequently encourages shifts in attention to new and more interesting targets.

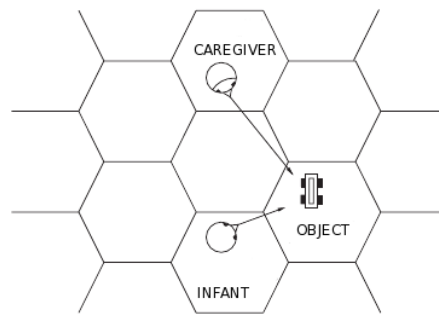


Figure 3.4: A visual overview of the gaze following model (adapted from Triesch *et al.*, 2006.)

The gaze following model is represented pictorially in Figure 3.4. The infant and caregiver are assumed to begin facing each other and to remain stationary (other than to shift gaze.) There are N possible discrete locations that can be the target of a gaze. The

simulation proceeds through discrete time steps that correspond to approximately $\frac{1}{4}$ second. Up to one gaze is possible for each time step and there is at least one interesting stimulus present in the environment at any time (defined as a ‘target’.) Each target has an associated minimum number of time steps that it remains stationary, after which it has a probability of shifting to a new random location. When the target moves, the caregiver will automatically shift its gaze to follow it.

At each time step, the infant must decide to maintain gaze on the current location or shift it to one of the others. The infant will receive a reward based on one of four possible classes of stimuli in their visual field: frontal view of caregiver; profile view of caregiver; target; or, nothing. The value of the current stimulus category is then attenuated by the current degree of habituation. Thus, the reward associated with a particular stimulus in a given location will decrease with time.

Over time, the modelled infant learns a few general heuristics for maximising reward. First, the habituation attenuation encourages exploratory gaze shifts over staring at one location for too long. Second, periodic gaze shifts back to the caregiver tend to be rewarding. Third, the caregiver’s gaze direction is a good predictor of reward. These principles were also demonstrated in simulations with multiple simultaneous targets, albeit at a slower learning rate.

The authors varied different parameters of the model to analyse their effect. The probability that the caregiver would be looking at the new location of a target proved to be very important. When the probability was too low, the correct gaze following behaviour did not emerge in the infant. This result highlights the importance of the social dimension to the development of gaze following (Triesch *et al.*, 2006). Different values of the learning rate showed that an intermediate value gave the best results, which is very common with neural networks. For the habituation parameter, higher values proved to be correlated with faster learning rates. When the habituation rate was set to zero, gaze following still emerged but it was learned at a much slower rate.

Triesch *et al.* also tweaked the components of their Basic Set to model special populations, such as children with autism spectrum disorders or what is known as Williams syndrome. Such children can be classified very loosely as hyposocial or *hypersocial*, respectively. These disorders were simulated by adjusting the reward value for looking at the caregiver relative to the reward for looking at the target, respectively decreasing or increasing it. When the relative value was small, gaze following developed abnormally and the infant spent much less time looking at the caregiver, consistent with an autistic child’s lack of interest in faces. When the relative value was larger, gaze

following was delayed, consistent with the shared attention deficits observed in children with Williams syndrome (as reviewed by Triesch *et al.*, 2006).

The gaze following model is very appealing both for its simplicity and its neurologically plausible learning scheme. Both of these characteristics are very important and I will revisit them in later chapters. The authors argue against the existence of an innate “joint attention module”, instead modelling how gaze following may emerge through a gradual learning process. Such a developmental trajectory also parallels many theories of language acquisition. Their disorder simulations are also noteworthy, demonstrating how slight changes in parameters can lead to abnormalities in the system. Of course, the gaze following model is not simulating language learning, so I will next discuss a model that uses social-pragmatic cues in such a context.

3.2.2 Yu and Ballard’s Unified Model

Yu and Ballard (2007) created a unified model of early word learning. The unified model incorporates statistical learning, as well as simulated social cues from visual and auditory information. One important feature of their model is that it was trained with information gleaned from video clips of real mother-child interactions.

The word learning situation is conceptualised as two streams (see also Siskind, 1996), one from audio and one from video. The video stream consisted of all the objects visible to the infant at any time, including the mother’s referents. The audio stream was constructed from all the words spoken by the mother. According to their methodology, each multi-word utterance and the corresponding visual scene constituted one learning situation. The authors describe the goal of each learning situation as building “several one-to-one mappings from many-to-many possible situations.” For example, the audio stream might contain the utterance “oh, there went the cow” and the visual stream might consist of a cow and a pig; any of the five words (or none of them) could refer to the pig or the cow. The one-to-one mappings that are learned will associate an individual word (i.e., a label that is heard) with a meaning or concept (i.e., the object seen) in what the authors refer to as a “word-to-world mapping”; essentially, the mapping problem.

The visual information is modelled using what is referred to as a joint attention “spotlight”. Whatever object the mother was gazing at is given extra weight in the mapping process, compared to any other objects in the infant’s visual field at that time. The auditory information is incorporated by analysing the low-level properties of the mother’s speech. Prosodically-distinctive words are considered more salient and

	statistical	prosodic	attention	prosodic-attention
precision	75%	78%	80%	83%
recall	58%	58%	73%	77%

Table 3.1: Summary of results from the unified model (Yu and Ballard, 2007).

also given extra weight in their algorithm.

A number of simulations were run using different elements of the unified model. In each case, measures of precision (percentage of learned mappings that were correct) and recall (percentage of correct mappings out of those that could possibly have been learned) were calculated from the results. The four conditions, summarised in Table 3.2.2, were as follows: statistical learning only (*statistical*); statistical learning with prosodic cues (*prosodic*); statistical learning with joint attention (*attention*); statistical learning with both prosodic cues and joint attention (*prosodic-attention*). It is important to note that the authors calculated the precision and recall of chance random mapping at 5.3% and 15.2%, respectively. It is clear from the results that all implementations were capable of word learning, but the ones that included social cues were more successful.

When Yu and Ballard reinterpreted their data in terms of association probabilities, the differences were even more striking (see Figure 3.5.) An association probability is the probability of a meaning given a word. The left bars in the figure represent the mean association probabilities for correctly-learned pairs. The right bars show the mean association probabilities for the first three incorrect pairs for each object.

Again, the results from statistical learning alone are acceptable. However, the advantage of adding in the social cues is clear. Simulating joint attention, especially, increases the probabilities of correct word learning and decreases the probabilities of incorrect mappings. As the authors point out, a significant difference between the two will allow for a threshold to be defined which could be used as the criterion for correctness.

The Unified Model is a very convincing simulation, but there are a couple of outstanding considerations. First, the network is similar to that of Siskind in that it takes a weighted set of semantic items and maps them to a set of words. There is neither a true concept of communicative intentions, nor a process for the child to learn the importance of such situations. I propose that the manner in which the streams are

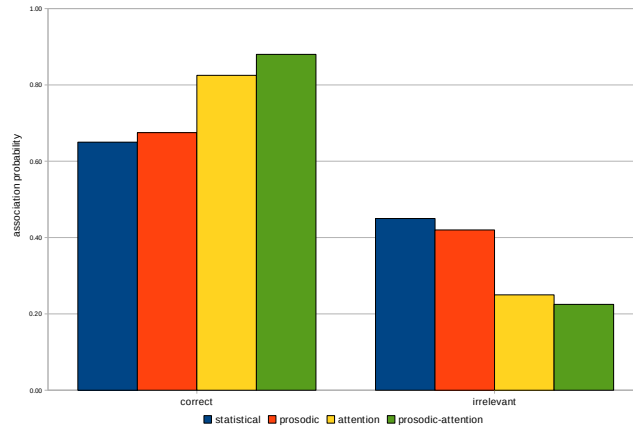


Figure 3.5: Comparison of association probabilities from the unified model. (Adapted from Yu and Ballard, 2007.)

partitioned does not possess the appropriate granularity for representing the individual events in a multi-word utterance, a point which I will elaborate upon in the next chapter. Finally, the neural substrates of the learning mechanism remain an open question.

3.3 Discussion

Each of the models I have described in this chapter produced interesting results. Statistical learning models—however well they realistically model statistical learning processes—are not fully convincing because they do not tell the whole story. I do not dispute the importance of pattern-recognition and focused attention to language learning, but I assert that the big picture must be taken into consideration. Infants are not locked away in isolation, mining masses of linguistic data for useful patterns (see also Butterworth and Harris, 1994, pp. 124–129). They are learning language in a very social context, and there is too much evidence regarding the influence of that environment to simply disregard it.

One model (Triesch *et al.*, 2006) takes into account the importance of social cues, but not in any language-related capacity. Only one model that I am aware of (Yu and Ballard, 2007) integrates both intention recognition cues and statistical learning in a word-learning context. In the remainder of this thesis, I plan to explore some of the same territory as the Unified Model, with my own computational models. In later

chapters, I will carry it a few steps further by considering the influence of intention recognition on word learning from a neuroscience perspective.

Chapter 4

Joint Attention as a Filter on Word Learning

The previous chapter reviewed existing computational models of word learning. In this thesis, I outline a new computational model that incorporates some of the better features of those other models, with the goal to simulate various aspects of word learning in social-communicative situations. First, I will summarise the background that motivated my design decisions. Then I will gradually explain my model, through a progressive series of experiments. I will suggest that the recognition of communicative intentions is a means to filter out irrelevant ambient speech, allowing an infant to focus on more valuable word learning situations.

4.1 Background

I will begin by conceptualising an infant’s environment as a set of parallel streams corresponding to different modalities.¹ One stream consists of auditory input, which bombards the infant from many sources: multiple human speakers, televisions, music, etc. In parallel, the infant’s focus of attention controls a separate, visual stream. This second stream consists of the sensorimotor information from the infant’s own actions and the observation of others’ actions, in the form of concepts. The correlation between the two modalities is quite low (Sabbagh and Baldwin, 2005); a word heard at a given moment will only occasionally correspond to what is currently in the infant’s focus of

¹I will assume—as many researchers do (Siskind, 1996; Yu and Ballard, 2007)—that the infant is capable of segmenting an auditory stream into individual words and has the visual ability to discern objects.

attention. A simple illustration is shown as a timeline in Table 4.1, where the two streams are actually congruent in only one case (i.e., seeing a cat and hearing the word *cat*²—see also Yu and Ballard, 2007). While cross-situational learning alone could eventually distinguish the real correspondence from those that occur by chance, that would be an inefficient way of learning. I propose that it would be more efficient for the child to learn how to identify congruent examples through the use of social-pragmatic cues, consistent with Tomasello’s theories.

Object seen	DOG	DOG	MOTHER	CAT	FOX	DOG	BABY
Word heard	<i>run</i>	<i>sleep</i>	<i>baby</i>	<i>cat</i>	<i>cat</i>	<i>jump</i>	<i>eat</i>

Table 4.1: The top row represents visual, sensorimotor information.

The bottom row shows examples of coinciding emphasised speech.

The single congruent case is highlighted.

Clearly, these two perceptual streams are conceptually similar to those used by other models discussed in Chapter 3 (Siskind, 1996; Yu and Ballard, 2007). However, the two other models attempted to map multi-word utterances to sets of concepts. My streams differ in their granularity by isolating individual concepts and single-word utterances. It would be more realistic to assume that the child hears *several* words at each step, especially in a noisy environment. For simplicity, however, I am assuming that the child pays special attention to certain emphasised words and these emphasised words are the ones that appear in my utterance stream.

What I have, then, is another representation of the mapping problem discussed earlier (e.g., Siskind, 1996; Yu and Ballard, 2007). The problem can be solved by picking out what is relevant and ignoring the irrelevant (see also Elman, 1993). The ideas of social-pragmatic theory and the evidence of the joint attention studies (e.g., Baldwin, 1993) help theorise about how an infant could accomplish such a goal. The model I will describe simulates a child who can identify communicative acts and respond to joint attention. The network’s goal is to filter noisy parallel streams of verbal and sensorimotor input that are only sporadically correlated, highlighting the correlated cases so they can be successfully processed by a word learning mechanism. It is assumed that this mechanism can be turned off at times that are inappropriate to successful mapping, preventing the infant’s vocabulary from becoming muddled by incorrect mappings.

²Hereafter I will use small capital letters to indicate CONCEPTS and SENSORIMOTOR REPRESENTATIONS and italics to represent *words and linguistic representations*.

4.2 Experiment A: The Simple Word Learning Model

4.2.1 Introduction

I will begin with an artificial neural network model of single-word learning³, which I will refer to as the **Simple Word Learning Model (SWLM)**. The SWLM learns a mapping between concepts and words from a “perfect” set of training inputs, without any of the noise shown in 4.1. It is an idealised version that, in terms of the example of Table 4.1, focuses only on the congruent cases like CAT-*cat* and ignores all other input. Therefore, it allows me to model the hypothesised word learning mechanism and assess how it learns when it is fed appropriate information only.

The implementation⁴ is a simple feed-forward network, as illustrated in Figure 4.1. The architecture used, known as a **multilayer perceptron**, is very common in the field of neural networks. An input vector is applied to the nodes in the input layer. Each node in the hidden layer receives a weighted sum of those inputs and translates it into a discrete value, based on some activation function. The values from that layer are projected to the output layer, where the weighted sum is again passed through an activation function. The outputs of the final nodes form an output vector that is used as the overall response of the network. The connections between the layers start with random weights and are strengthened or weakened during training. Weights are adjusted according to the standard backpropagation algorithm (Haykin, 1999), which is why the architecture is also known as a **backprop network**. Backpropagation is a form of supervised learning, wherein the output at each step is compared to the expected result (i.e., the correct answer) and any difference between the two becomes an error signal which is used to assign credit or blame.

4.2.2 Methods

The input to the network is a sensorimotor concept (e.g., DOG.) These are represented using a one-hot coding scheme; i.e., if DOG is one of three possible concepts then it might be encoded as 100, meaning the input to the first node would be turned on (see

³Although 18- to 24-month-olds are beginning to produce multi-word utterances (Tomasello, 2003, p. 113), I will concentrate on the learning of single words.

⁴All neural networks were implemented using the Java programming language (Sun Microsystems, Santa Clara, CA, USA), JDK 1.5.0.14 for Linux RedHat and 1.6.0.05 for Windows. Simulations and test set generation were generally run using the Eclipse integrated development environment (Eclipse Foundation, Ottawa, Canada) version 3.3.0 (Linux) or 3.2.0 (Windows).

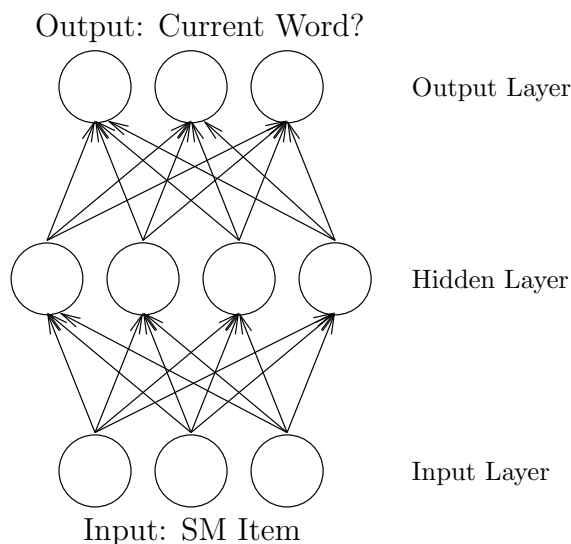


Figure 4.1: A reduced representation of the Simple Word Learning Model, implemented as a backpropagation neural network with one hidden layer.

also Elman, 1993). Thus, each input is independent, with no embedded semantic information. Activation propagates forward and the on-off statuses of the output nodes are translated into a representation of the word label. After training, the connections feeding forward between the layers produce an output which provides a one-hot encoding of the correct word label (e.g., *dog* might be 001) for each given input.

These simulations are intended as a proof-of-concept, both to reproduce some previous results (van Oijen, 2006) and to demonstrate that the proposed architecture is capable of learning the task at hand. By 24 months, the average child is expected to have acquired a vocabulary of 300 words (as cited by Ganger and Brent, 2004). The proportion of these that represent object names or other nouns is a topic of some debate, ranging from around 40% (Bloom, 2000) up to approximately 70% (Ganger and Brent, 2004). For the sake of these simulations, I will assume that a vocabulary of 150 nouns is sufficient to model a 24-month-old.⁵

A very simple training set was created, containing one instance of each concept (the input) matched with its corresponding word (the desired, or teaching output.) One training epoch consists of one run through the training set. As each input is presented, activation propagates through the network and an error term is generated based on the difference between actual and expected output; based on this term, weight

⁵I will explain why the network only needs to learn nouns, specifically, in Chapter 5.

adjustments are calculated and stored in memory, but not yet applied. At the end of each epoch, the cumulative changes across all training examples are applied to the weights. This approach is known as ‘batch learning’. After every 100 epochs, the network is probed to see how many of the possible inputs would produce the desired output. The number of words learned up until that point is defined as the number of ‘correct’ input-output mappings produced. Training continues until a vocabulary size goal is reached, meaning that the network would produce the correct word output for 100% of the input concepts.

The nodes implement a sigmoidal activation function for all simulations. Because of the way the input patterns are encoded, one input neuron is required for each concept to be learned and one output neuron is required for each possible word. There have been metrics proposed for choosing an appropriate number of hidden neurons in perceptrons, but the process is almost more of an art than a science. A problem of this sort, with fully-orthogonal inputs and no noise, is simple enough that a high number of hidden nodes is not required. One such guideline is not to have more than twice as many hidden units as input units (Swingler, 1996, p. 55). Because the smallest number of inputs I test is 10, I decided to use 20 units (i.e., $10 * 2$) in the hidden layer for each simulation. The other parameters used are a learning constant of 0.25 and momentum of 0.9, which are common initial parameters for a backprop network (Swingler, 1996, p. 65).

4.2.3 Results

Table 4.2 summarises results with the SWLM, showing sample training times with 20 hidden units, for different vocabulary sizes.

Vocabulary Size (words)	Trainings Epochs required for 100% accuracy
10	3 800
25	67 100
50	242 900
100	1 064 600
150	1 357 400

Table 4.2: SWLM training results for different problem sizes, showing the number of epochs required to map the vocabulary with 100% accuracy.

These results are intended as a guideline only, as each represents an n of only 1 and there will be a great deal of variability in training time. However, as would be expected, the results show a clear trend that the number of training epochs required increases with increasing vocabulary size.

4.2.4 Discussion

I have demonstrated that the backprop architecture is capable of scaling up to a vocabulary size with realistic proportions for a 24-month-old. Training at the full size does require a large number of epochs for the network to reach 100% accuracy, however. In the interests of expediency, I will run future experiments on a smaller vocabulary and generalise the results to reach conclusions about performance.

The SWLM works very well for learning isolated words. After training, it essentially becomes a word predictor (see also Elman, 1993); given a concept as input, it produces the corresponding lexical item as output. It also satisfies the difficult bootstrapping constraint of language-learning systems, learning with no prior knowledge.

However, initial word learning is not the focus of my research, and the SWLM cannot accomplish all of my goals. First of all, the network was trained on very artificial input sets that are not cluttered with the inherent noise of ambient speech. Furthermore, it does not factor in any of the social cues that supposedly facilitate the process of word learning. The next experiment will improve upon the SWLM, adding a model of joint attention in order to simulate the benefit of recognising such cues.

4.3 Experiment B: Word Learning with a Joint Attention Filter

4.3.1 Introduction

Consistent with the developmental psychology literature, I propose a joint attention model as the means to filter noise from ambient speech. The hypothesis is that a high level of noise will severely impair the learning of the SWLM. An adapted version, filtered and with less noise, should learn faster. Such a model would suggest that recognising and focusing on communicative situations of joint attention will facilitate word acquisition.

Of course, basic joint attention is not the only requirement. Even if a child consis-

tently follows the referential intent (e.g., gaze) of the mother, there will not always be a word-learning situation because she might not actually be speaking at the time. What is crucial is that she also says a word that refers to what she is looking at; following her gaze to the correct target will only then allow the infant to form a sensorimotor representation of her referent and be able to map the concept seen to the word heard. This idea is illustrated in the timelines of Table 4.3. The table assumes a simple form of ‘whole event’ recognition, because the cue for a communicative situation is a complete event or proposition: seeing the mother and seeing that she is talking. Thus, step 1 depicts a dog sleeping, step 2 depicts a dog jumping, and so on.⁶ Note that the timeline uses the placeholder `_NO_ACTION_` to represent attending to an agent, with no associated action (e.g., attending to a dog that is not engaged in any particular action.) The model assumes that an action cannot be witnessed independent of some sort of agent. The highlighted steps identify the target case of an ideal word learning situation, in which the word heard will be correlated with a referent that is present in the environment. In general, it is assumed that the child is exploring his environment by looking at salient stimuli. When the mother begins talking (as in step 3), a communicative intention is recognised. At this point, the child follows the mother’s gaze to the object of her attention. The next time step, 4, reflects the establishment of joint attention with the mother. The word heard is associated with the visual information observed at the same time step. Note the implicit connection between the event at steps 3 and 4: because of attention following, the concept at 4 represents the *content* of the speech event at step 3.

	1	2	3	4	5
Object	DOG	DOG	MOTHER	CAT	FOX
Action	SLEEP	JUMP	TALK	RUN	_NO_ACTION_
Word	<i>run</i>	<i>sleep</i>	<i>dog</i>	<i>cat</i>	<i>cat</i>

Table 4.3: Illustration of the joint attention model. The top two rows display visual information. The bottom row shows the contiguous word held in the phonological loop. The crucial trigger is MOTHER-TALK, and the subsequent learning opportunity is highlighted.

⁶Objects and actions are kept separate in these event separations to reflect the fact that they are recognised by different perceptual pathways, in anticipation of an extension that I will present in Chapter 5.

At this point, I would like to clarify the contents of Table 4.3, especially with respect to the MOTHER-TALK trigger. My proposal is that the word heard at a given time step consists of the contents of the **phonological loop (PL)**. The phonological loop is a theoretical subcomponent of working memory (WM). It stores acoustic memory traces that, unless rehearsed subvocally, fade after about two seconds (Baddeley and Hitch, 1994). In other words, the time steps above consist of the word that is in the PL (i.e., the one just heard) and the representations of objects and actions currently in the focus of attention. At step 3, the infant realises that the mother is talking and forms a representation of MOTHER-TALK. However, *dog* is not what the mother is saying; rather, *dog* is stored in the phonological loop from some other auditory source—assume a crowded room or, perhaps, a television. The PL gives the child a chance to do some learning even after the mother’s words have been spoken. Since recognising that the mother is talking requires hearing her words, and since it takes time to establish joint attention, a buffer for the phonological input is a necessary component of any theory which ascribes a role to social-pragmatic cues. It is not until step 4 that the infant processes the word *cat*, which is what the mother is actually saying.

In terms of the model, it should be possible to learn to recognise that MOTHER-TALK has special significance for word learning, and to exploit this knowledge to establish joint attention and ‘turn on’ the word learning mechanism when this event occurs. Using this approach, a noisy simulation of ambient speech can be filtered by only allowing training input through to the SWLM at these appropriate times. According to this strategy, language learning is only enabled when the crucial trigger is identified. When the infant recognises a MOTHER-TALK situation, he follows her gaze to see what she is talking about. At this point in the model, language learning is enabled and the SWLM network is trained to map the spoken label onto the attended-to sensorimotor information. (E.g., (a) Mother says *cat* and indicates a cat; (b) baby looks at the cat and hears *cat*.) My **joint attention filter (JAF)**⁷ is reminiscent of the ‘joint attention spotlight’ (Yu and Ballard, 2007), which gave extra weight to training items that were identified by social cues. However, it also explicitly models the role of recognising and representing individual communicative intentions.

Work by Harris *et al.* (1986) provides a starting point for setting realistic parameters. The study consisted of videotaping mothers interacting with 16-month-olds and then assessing the language development of the same infants at 24 months of age. A

⁷More accurately, it could be described as a ‘joint attention / communicative intention recognition filter’, as its purpose is more specific than identifying general cases of joint attention.

correlation was found between the number of references that mothers made to objects in the child’s current focus of attention at 16 months and the child’s later rate of language development. Ten children were classified into ‘normal’ and ‘slow’ groups, according to their language development. Table 4.4 shows data of the authors that I have used to estimate parameters for my experiments.

	Mother-initiated	Referential	Current	Specific	Result
Normal	0.31	0.96	0.74	0.46	0.10
Slow	0.27	0.85	0.49	0.25	0.03
Mean	0.29	0.91	0.62	0.36	0.06

Table 4.4: Estimating the percentage of communicative events involving joint attention (based on Harris *et al.*, 1986).

The first column is the mean proportion of verbal episodes that were mother-initiated. The second column is the proportion of mother-initiated episodes that I have classified as referential (i.e., involving gaze and/or action on the part of the mother), based on the methods of Harris *et al.*. The *Current* column is the proportion of the mother’s utterances for which the child was focused on the actual referent at the time. Harris *et al.* reported these proportions in categories that did not seem to be mutually exclusive, so I have picked the largest proportion in each case. The *Specific* column shows the proportion of utterances to current referents that were specific (e.g., “ball” rather than “thing”.) The final column is the product of the other columns, giving a rough approximation of the percentage of instances of joint attention. This value ranges from 0.03 for the slow group to 0.10 for the normal group, with a mean of approximately 0.06. Even if all mother-initiated communications qualified, the proportion would still be much less than 0.40, based on the highest value in column one. In general, these estimates are likely to be inflated for a number of reasons. First, I chose the category with the maximum proportion of current referents in each case. Second, the experimental procedure was highly artificial as children were observed in a concentrated play session of 10 minutes. No other sources of auditory input were present, suggesting a much smaller proportion of uncorrelated overheard speech than an infant would likely encounter in everyday life. Finally, the proportion might not generalise well to other cultures where CDS might be used less often or take a different form (Fernald and Morikawa, 1993; Floor and Akhtar, 2006). However, if my hypothesis holds for a value of 0.10, then it should be at least as applicable to any real world

examples with smaller proportions.

4.3.2 Methods

A network with the JAF included is represented in Figure 4.2. The thickness of the connection lines indicates the relative weight strengths that are hard-coded for the current experiment (see the next Chapter for a discussion of how these weights could be trained.) Object-action pairs are presented at the inputs (the working memory (WM) layer.) Each input is fully-connected to the two nodes in the gating layer (GL), feed-forward. One node in this middle layer (labelled with “Off” in the figure) represents a simple state of observation and, when activated, inhibits the SWLM so no word learning is attempted. When the other node in the middle layer (“On” in the figure) is activated, it enables the SWLM. Activating this node also allows the sensorimotor information to pass through and be used as input to the SWLM. Therefore, this toggle is the implementation of my proposal that the infant can switch into and out of a word learning mode.

During training, the associated word term is used as teaching output by comparing it to the actual output of the SWLM (as in Section 4.2.2.) The filter simulates a child who has perfected the ability to recognise a communicative situation (i.e., MOTHER-TALK); it is hard-coded to correctly identify all relevant speech. It is also assumed that the child has mastered the ability to respond to joint attention, knowing—and being reliably able—to follow the mother’s gaze to what is occupying her attention. Thus, the filter will always activate and inhibit the SWLM at the appropriate times.

The auditory and visual streams are simulated by a series of agent-action-word trios. In each case, the word can correlate with the agent (noun), with the action (verb), or be completely uncorrelated (e.g., ELEPHANT-SLEEP-*eat*; DOG-SMILE-*elephant*; CAT-SLEEP-*cat*; ELEPHANT-SMILE-*smile*; etc.) The trios were randomly generated from the input set of possible words, using the following constraints and parameters:

1. A child-directed speech (CDS) parameter, the probability that a given input would represent a joint attention situation, was set to 0.1. Thus, approximately one in ten trios was MOTHER-TALK-*[word]*, followed by a correlated trio that represents a valid learning situation.
2. Generation of input trios continued until each possible noun and verb had been represented in at least one such joint attention situation.

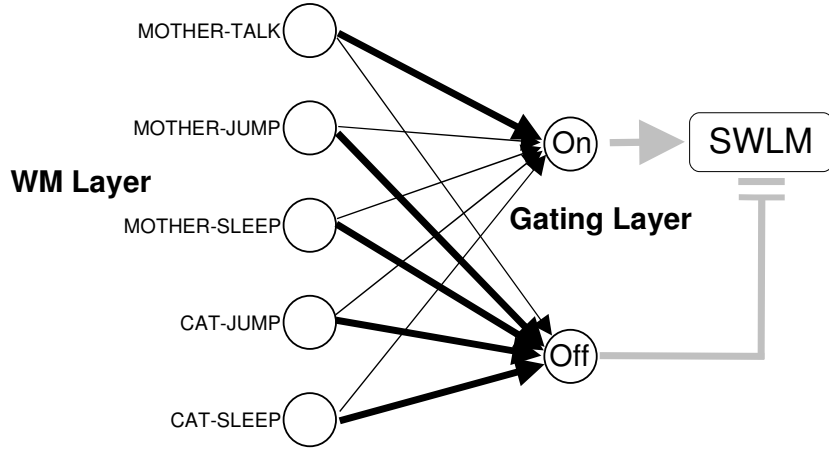


Figure 4.2: An extension to the SWLM. The original neural network (Figure 4.1) is gated by a preceding network that implements the joint attention filter. Thicker lines between the layers indicate stronger connections.

The goal of each simulation is to learn the correct mapping between the object concepts and their corresponding words. (Verbs and other word classes are not learned in this experiment; see Chapter 5.) The training set is first used in a series of statistical learning (SL) training runs without a filter, effectively reducing the network structure to that of the SWLM. For the second type, joint attention (JA) training runs, the filter is activated to simulate joint attention. The two types of runs can therefore be compared to analyse the effect of the JAF on learning speed.

The vocabulary is comprised of 25 nouns and 5 verbs. Training continues until the network can produce a 100% correct mapping for all nouns. The SWLM used in the previous experiment was modified slightly because of the extra noise that would be added the training set. The number of hidden nodes (50) was increased to be double the number of inputs (25). The learning constant was also reduced to 0.4 to avoid over-fitting to ‘bad’ training examples from uncorrelated input trios. These values may seem arbitrary, but trial and error demonstrated that they are effective. Although both of the altered parameters would likely affect training speed, possible confounds are avoided by keeping them constant across the experiment. Training is repeated ten times with and ten times without the JAF, using a new network each time. Thus, 20 result sets (2 filter conditions * 10 training runs) are generated for the one training set.

4.3.3 Results

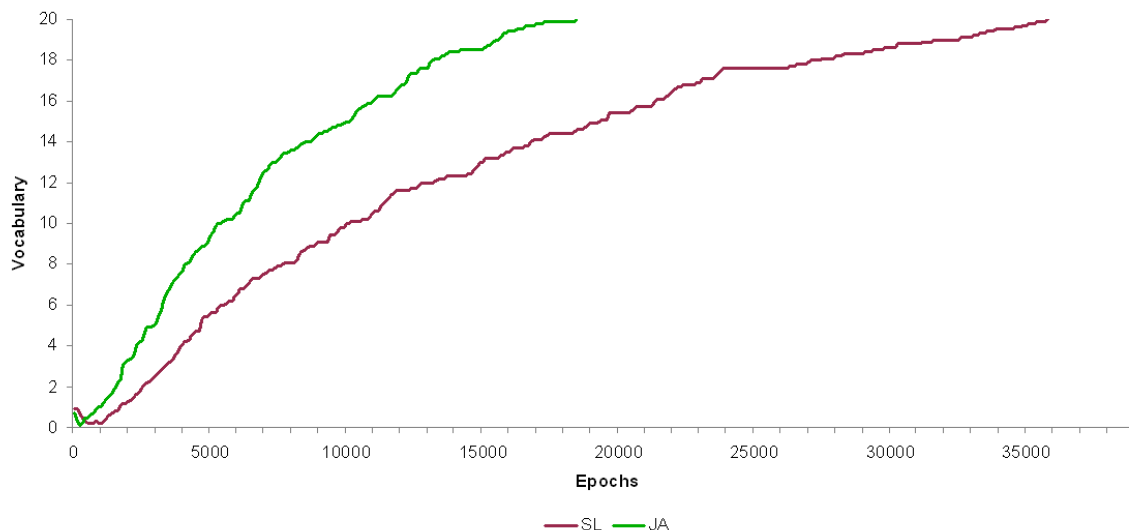


Figure 4.3: Experiment B. A comparison of learning speed with (*JA*) and without (*SL*) joint attention filtering.

Success is measured by the number of words correctly learned after a given number of training epochs. Data were collapsed across each set of ten training runs by taking the mean of the total vocabulary at any 100-epoch milestone, yielding two data sets (filtered and unfiltered.) Results across the first twenty words are shown in Figure 4.3.

The JAF had a clear effect, causing the network to learn the vocabulary in fewer epochs. Results are summarised in Figure 4.4. A one-tailed, paired t-test⁸—of the total number of epochs each network required to learn the vocabulary—confirmed that learning was significantly faster with the JAF than without ($p < 0.01$).

4.3.4 Discussion

The purpose of this experiment was to examine the performance of a network which is hard-coded to respond to a communicative event by enabling a word-learning network. The network was able to learn the vocabulary with or without the assistance of joint attention. However, in simulations with the JAF, the network was able to learn the full set of 20 nouns much faster than it did with the unfiltered input. The reason I only present the first 20 nouns learned is that the learning curves tend to asymptote as the

⁸Statistical analyses were calculated using Excel 2002 SP3 (Microsoft, Redmond, WA, USA.)

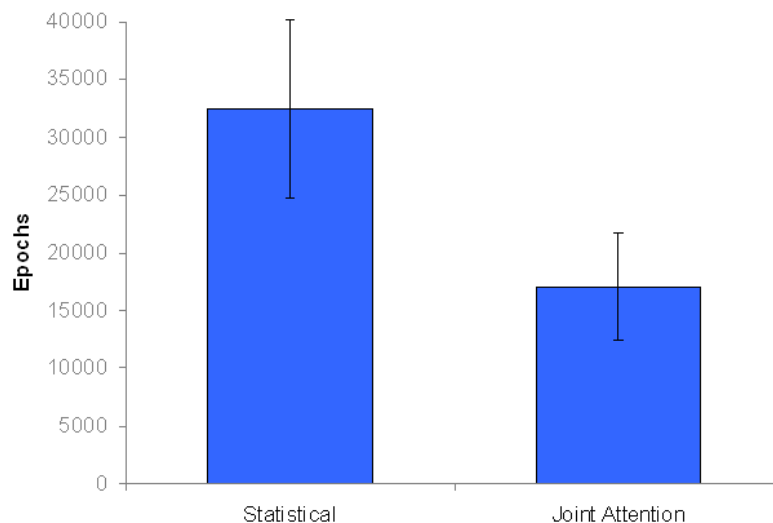


Figure 4.4: Mean number of epochs required to learn the 20 words in Experiment B, with and without the JAF. Error bars represent one standard deviation.

network approaches the point where it has learned the complete vocabulary. A real infant would not be expected to show a ceiling effect in his rate of word acquisition—especially at such a small vocabulary size.

This experiment is intended as a proof-of-concept that word acquisition is facilitated by deploying joint attentional skills when communicative situations are recognised. Of course, there were a great number of simplifications and abstractions used in the model above. For future experiments, I will slowly tease out some of these constraints to strengthen my argument. First, I will consider the point at which the JAF is introduced. For this experiment, the filter was enabled at the beginning of word acquisition. As discussed in Chapter 2, word production begins around 12 months of age but responding to joint attention is not mastered until around 18 months. The next experiment will look at the effect of introducing the JAF after word learning has already begun.

4.4 Experiment C: Development of Joint Attention

4.4.1 Introduction

As already discussed, an infant will be able to robustly respond to joint attention at about 18 months. According to the literature (Bloom, 2000; Ganger and Brent, 2004; O’Grady, 2005), he would have a vocabulary of approximately 50 words at this point. This fact does not contradict my hypothesis that joint attention speeds up word learning, but it does imply that it is not essential for early word learning. For this next experiment, I will allow the network to learn a few concept-word meanings before the JAF is activated.

The previous experiment also did not allow for those cases where the child will overhear a word that just happens to correlate with what his attention is currently focused on (i.e., he is looking at a dog at the same time another person says *dog*), without any sort of joint attention taking place. After all, an infant will experience many different sorts of learning situations. I will refer to such events as **chance correlation**. This parameter can also be very roughly estimated from Harris *et al.*’s experiment. Their data, as presented, make it difficult to reach a definitive answer. However, if we take all cases across the two groups where the mother is *not* initiating an episode with referential intent ($1 - 0.29 * 0.91$), the maximum proportion of those that refer to a current referent (0.62), and the proportion of those that refer to a specific object (0.36), we arrive at approximately 0.16. Again, if anything, this number is probably an overestimate.

This experiment is based on previous work (Caza, 2007), but presents new data.

4.4.2 Methods

Training sets were generated using the methods above, with the addition of a chance correlation (CC). This parameter was implemented as a probability that correct correlations (e.g., DOG-NO-VERB-*dog*) would have a minimal occurrence even outside of joint attention situations. This parameter represents other word learning situations that the infant might experience. These correlations are over and above any that would otherwise occur during random generation (i.e., with a vocabulary of 30 words, each word already has about a 1 in 30 chance of occurring regardless of any of the other parameters of the system.) To reduce possible confounds, the CC parameter was fixed for the entire experiment at a value of 0.15, close to the rough estimation of 0.16 calcu-

lated above. Note that an artificially high CC value should only make statistical-only learning faster and would, if anything, reduce the apparent effectiveness of the JAF by making the differences in learning rates seem smaller.

As before, simulations are run again with and without the JAF. For the former type, the cognitive development of joint attention is simulated by enabling the filter after a subset of the nouns has been successfully learned. Once enabled, the gate is fully functional for the remainder of the training period. For such runs, the filter is activated after the network can map at least five (any five) of the nouns with 100% accuracy. This parameter is based on the infant learning to respond to joint attention around 18 months of age (Triesch *et al.*, 2006; Baldwin, 1991), at which point they have already learned around 50 words of their 24-month-old vocabulary of 300 words (i.e., between 10 and 20%).⁹

Although Harris *et al.* (1986) provided a starting point, it is difficult to assess what percentages of actual speech heard by an infant would fall into the categories of uncorrelated, correlated by chance, and correlated by joint attention. As such, a variety of CDS values are used in this experiment in order to assess the effect the parameter has on learning. Simulations are run using different training sets generated using the following values of CDS: 0.10, 0.20, 0.60, and 0.80. Training is repeated ten times for each training set with and ten times without the JAF, using a new network each time. Thus, 80 result sets (2 filter conditions * 10 training runs * 4 training sets) were generated in total.

4.4.3 Results

Data were collapsed across each set of ten training runs by taking the mean of the total vocabulary at any time point, yielding two data sets (filtered and unfiltered) for each parameter set. Results are shown in Figures 4.5 through 4.8. The point at which the JAF was enabled, where applicable, is also shown.

The JAF had a clear effect for each value of CDS, as before, increasing the rate of word learning. Results are summarised in Figure 4.9. A two-factor analysis of variance (ANOVA) showed a main effect of filter and of CDS, as well as an interaction between them. A single factor ANOVA, collapsed across filter conditions, showed no significant difference between values of CDS ($p > 0.17$). Collapsing across all values of CDS, confirmed that the network learned significantly faster with the JAF ($p <$

⁹Since Experiment A proved that the SWLM architecture could learn up to 150 nouns, a scaled-down vocabulary is used here for demonstration purposes.

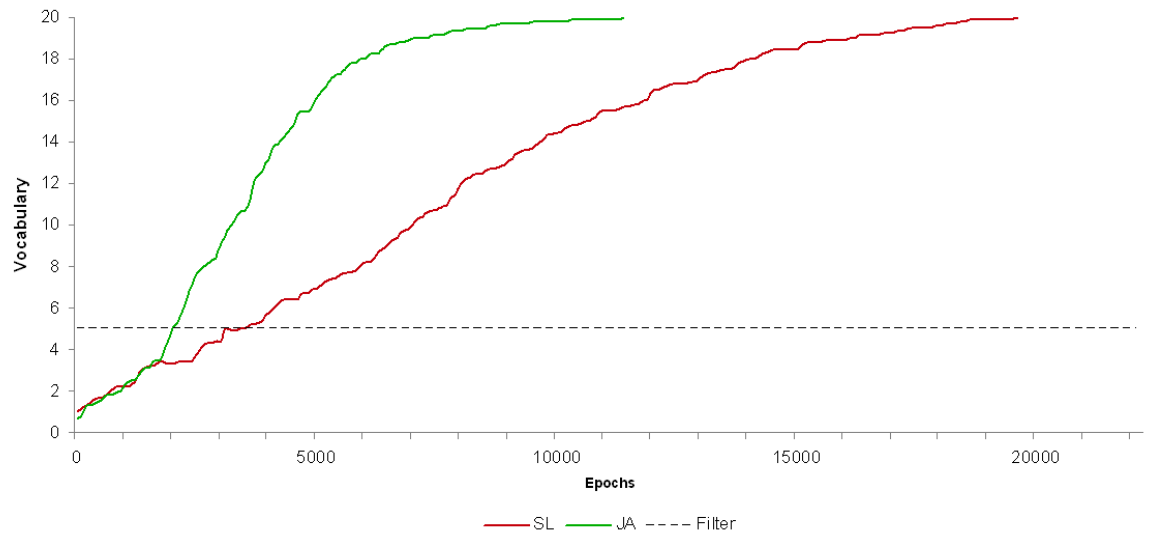


Figure 4.5: Experiment C1. Results with CDS equal to 0.10.

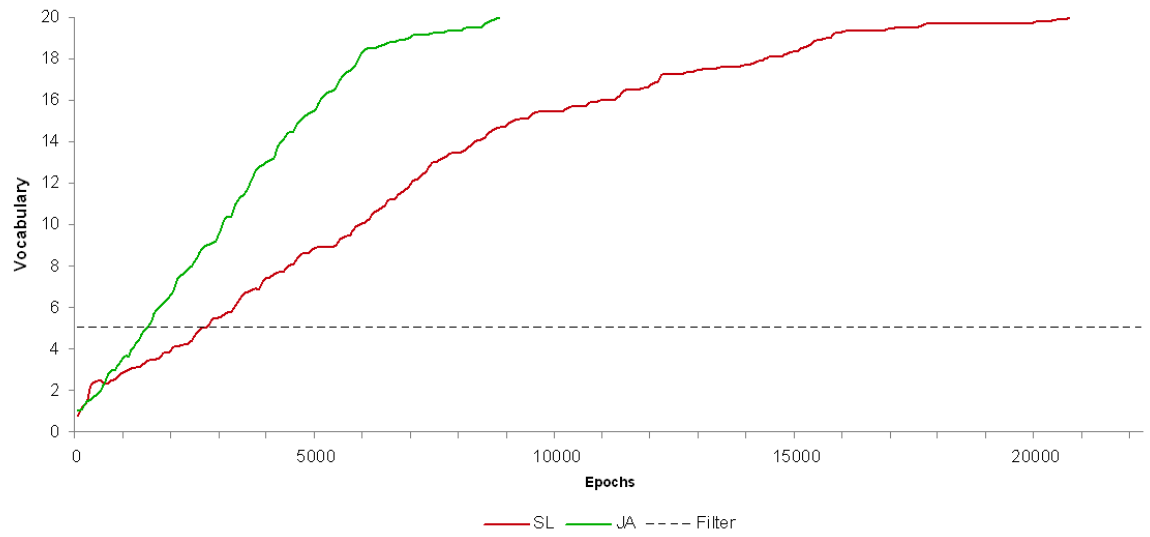


Figure 4.6: Experiment C2. Results with CDS equal to 0.20

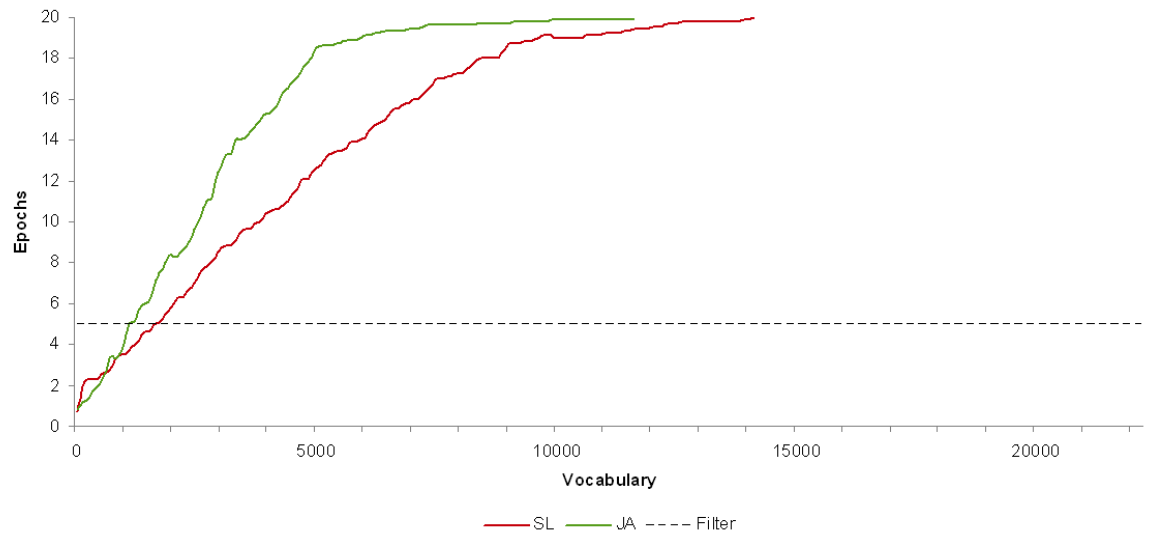


Figure 4.7: Experiment C3. Results with CDS equal to 0.60.

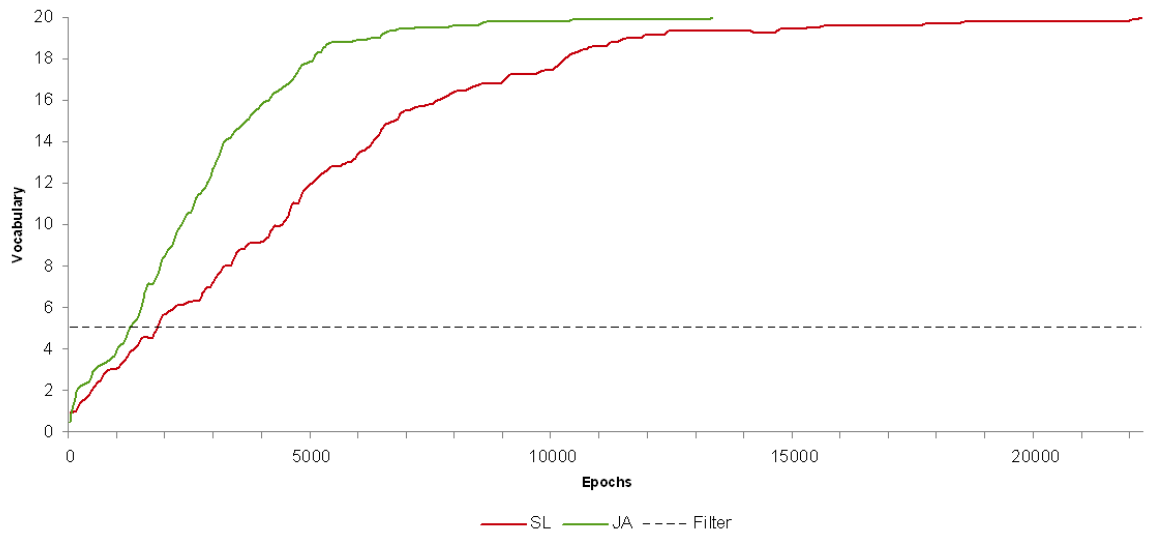


Figure 4.8: Experiment C4. Results with CDS equal to 0.80.

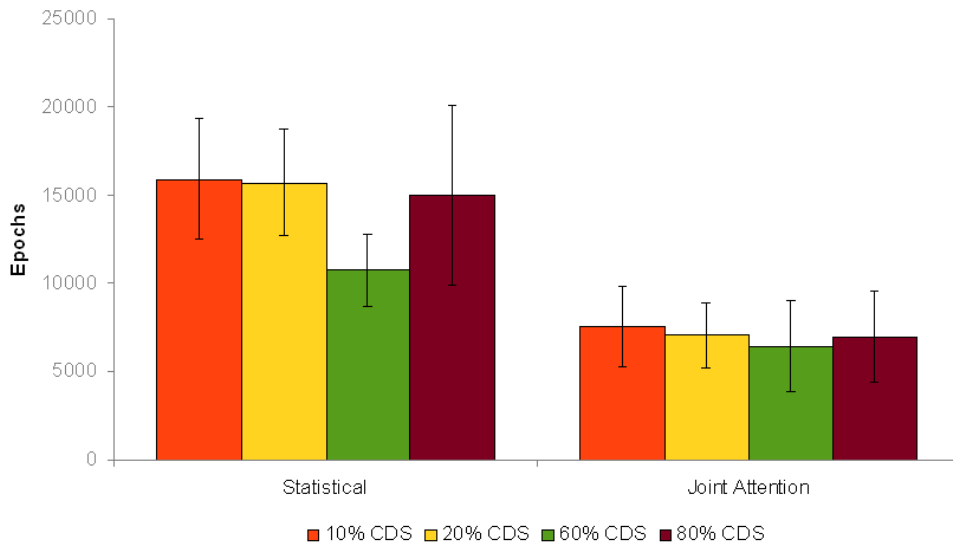


Figure 4.9: Mean number of epochs required to learn the vocabulary in Experiment C, for different values of CDS. Error bars represent one standard deviation.

0.01). It is very encouraging that the effect of the JAF is significant, even when a high proportion of the input is correlated. Based on Harris *et al.* (1986), however, values of CDS above 0.40 are highly improbable. Therefore, further analyses will only look at the two smaller values. A single-factor ANOVA of this population, CDS=0.10 or CDS=0.20, also showed a significant effect of the JAF filter ($p < 0.01$).

Further analysis was performed to ensure that the differences in learning rate were not caused by initial weights or some other characteristic of the different neural networks. Figure 4.10 summarises the learning times, collapsed across the two smaller values of CDS. Prior to any filter being activated, each network is essentially performing the same task, learning the first five words. There were no significant differences between networks with and without the filter, prior to it being activated ($p > 0.12$).

4.4.4 Discussion

This experiment has demonstrated that the JAF will cause acceleration even when it is activated after the word learning process has begun, just as it did when it was in place from the beginning.

It is interesting to consider whether any of these data show evidence of a vocabulary spurt. I do not intend to fully apply the rigorous statistical analyses of Ganger and

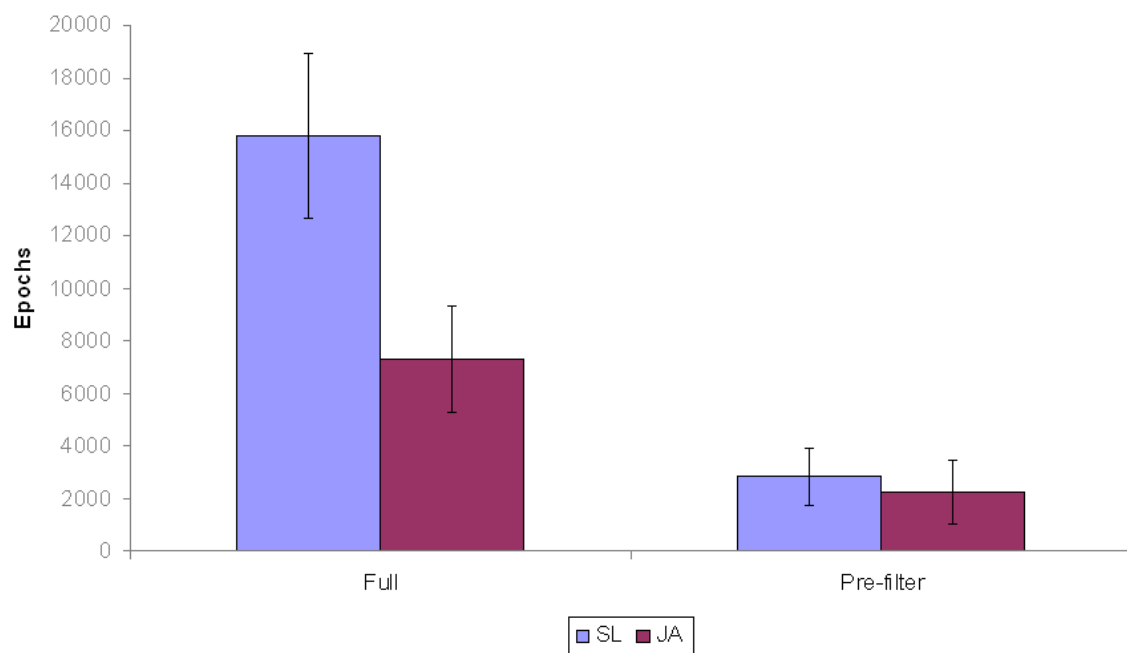


Figure 4.10: Mean number of epochs in Experiment C, for the smaller values of CDS, to learn all twenty words (*Full*) or the five before the filter is activated (*Pre-filter*).

Brent (2004) at this time. However, I can borrow from their methodology. They looked for spurts when the infant had a vocabulary between 20 and 90 words. Consistent with the 1:10 scale used in my experiment, that would translate to between 2 and 9 words. I will focus on the results of the training runs with CDS 0.10. Examination of Figure 4.5 shows no obvious point of inflection in the mean results, but this would be consistent with the evidence that many children do not demonstrate a spurt.

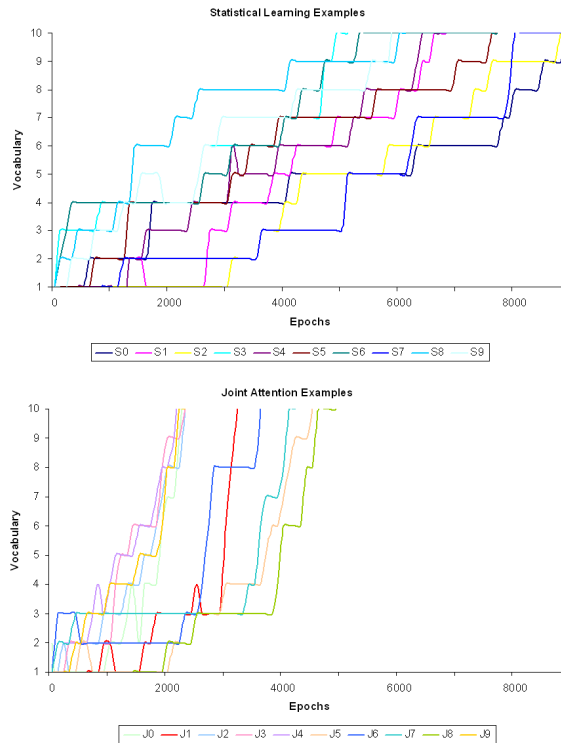


Figure 4.11: Examples of learning with CDS 0.10, from vocabulary size 1 to 10. (i) Statistical learning only; (ii) learning with joint attention.

The individual runs are plotted in Figure 4.11, for when the vocabulary was between 1 and 10 words. The statistical learning runs all advance in a fairly consistent stepwise fashion. The joint attention runs all show a fairly steep incline. However, the five on the left are consistently climbing throughout the 1 to 10 word period, without any obvious point of inflection. Tentatively, the five JA runs on the right *might* demonstrate acceleration in word learning. It would be premature to draw too many conclusions based on this qualitative analysis. However, my data do suggest two things: the network does not show a vocabulary spurt in all cases; and, the network will *only* show a vocabulary spurt in cases with the JAF.

4.5 General Discussion

While the evidence presented in this chapter is an excellent beginning, there are several unanswered questions. First, what is the mechanism which learns to turn on the JAF? The ability to recognise and take advantage of a communicative situation is not innate. A more interesting question is how an infant learns *when* to learn about words. How does he identify MOTHER-TALK as a tool and use it to facilitate future word learning? My model, as presented so far, only learns nouns. Verbs were only represented as a component of the MOTHER-TALK event. Finally, what is the neuroscientific basis for these abilities? These are some of the questions I will address in the next chapter.

Chapter 5

A Neurobiologically-Inspired Model of Intention Recognition and Word Learning

My joint attention word learning model, presented in the previous chapter, demonstrated the effectiveness of a device for selectively enabling word learning during communicative situations. In this chapter, I make further incremental improvements to the model. First, I turn to considerations of cognitive neuroscience. In Sections 5.1–5.3, I will review evidence for the neural correlates for object and action concepts, planning, and intention recognition. It is increasingly important for models of language processing to be expressed within a wider sensorimotor model, so these sections will be quite detailed. I will use this information to speculate about the brain regions involved in the various mechanisms of my model.

In Sections 5.4 and 5.5, I will re-express my model with an architecture that is more neurologically plausible. In Sections 5.6 and 5.7, I present an advanced model, involving a mechanism that learns to enable the word learning network in communicative situations.

5.1 Neural Substrates of Noun and Verb Concepts

The experiments run in the last chapter were focused on the learning of nouns. However, infants are not merely acquiring object labels. They also are gaining insight into actions and what to call them. Remember that the mapping problem (Siskind, 1996) classifies word learning as the process of forming associations between concepts and

spoken utterances. Before I extend my network to map both object and action concepts, I shall briefly discuss how these concepts are represented in the brain. (Refer to Figure 5.1 for a quick overview of the relevant brain regions.)

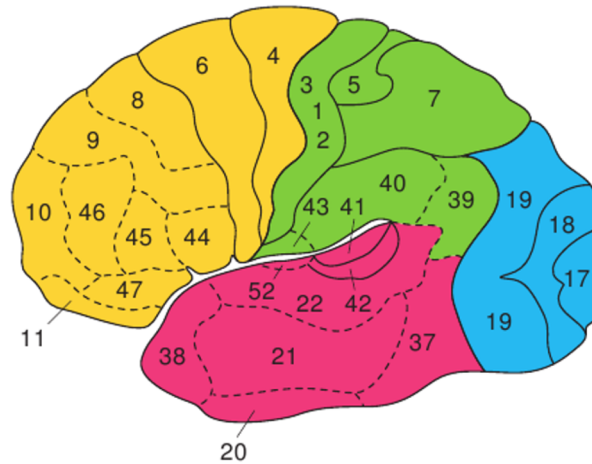


Figure 5.1: A lateral view of the major lobes of the human brain: frontal (yellow), parietal (green), occipital (blue), and temporal (red) (adapted from Gathercole, 1999). Numbers indicate Brodmann areas, with the prefrontal cortex (PFC) comprised of 8, 9, 10, 11, 44, 45, 46, and 47. Regions of interest in the PFC include dorsolateral (9 and 46), orbitofrontal (10 and 11), and Broca’s area (44 and 45). Other relevant locations include the premotor cortex (6), medial temporal gyrus (21), inferior temporal gyrus (20), and Wernicke’s area (posterior 22).

Gallese and Lakoff (2005) argue that there are ‘functional clusters’ in the brain, including parietal-premotor networks, with the characteristics required to represent concepts. They cite studies of macaque monkeys showing that premotor neurons in area F4 integrate the motor, somatosensory, visual, and auditory information used to control actions. Such neurons also respond when the monkey sees an object to which a specific action could be directed, suggesting that they are involved in forming an integrated representation of the concept behind the action. The authors also make a convincing case from the perspective of parsimony. The premotor region must be able to represent actions in some abstract sense. If the conceptual system were completely distinct, then it would need to duplicate some of the function of the premotor neurons, which seems like a waste of real estate. What this evidence suggests, therefore, is an overlap

between the neural systems in the monkey brain for controlling and conceptualising actions. Gallese and Lakoff extend their argument to human brains as well, due in part to evidence for analogous brain regions—the ventral premotor cortex (PMC) and intraparietal sulcus. Neuroimaging has found activity in the human ventral PMC when subjects observe tools, the same region that shows activity when those tools are used to perform an action (as cited by Gallese and Lakoff, 2005). Other compelling evidence comes from the discovery of what are called ‘mirror neurons’, which I will introduce in detail later.

Gallese and Lakoff also suggest that some of the neural correlates for language, as a form of conceptual representation, have a great deal of overlap with these same systems in the human brain. Again, I will explore this idea in more detail in the next section. For now, consider the possibility that mapping the sensorimotor concept of an action to the corresponding word, essentially accessing the semantics of the verb for that action, might activate some of the same brain regions as performing that action. To explore this idea, I will next review evidence from the experimental literature regarding the brain regions involved in retrieving the semantics of action and object concepts.

Damasio and Tranel (1993) published a very influential paper that suggested a double dissociation between the semantic systems for concrete nouns and verbs. A behavioural study contrasted three case-study patients with different lesions, against ten normal controls. For noun retrieval, subjects were presented with line drawings and asked to produce the most specific target word (e.g., “banana” rather than “fruit”). In the verb retrieval case, they were shown line drawings or photographs and asked to indicate what was happening. The two subjects who were impaired at noun retrieval had lesions in common areas that included medial temporal (MT) and inferior temporal (IT) gyri. The one subject who was impaired at verb retrieval had a lesion to the left posterior frontal region (LPFC) and PMC. A later study (Damasio, Grabowski, Tranel, *et al.*, 1996) examined 127 brain-lesioned adults and used positron emission tomography (PET) to scan 9 normal controls, producing corroborating results regarding noun retrieval depending on the temporal lobe. Further evidence for the double dissociation was found using case studies of other brain-damaged subjects (Daniele, Giustolisi, Silveri, *et al.*, 1994).

Recent studies using functional magnetic resonance imaging (fMRI) (e.g., Eichenbaum, Yonelinas, and Ranganath, 2007; Diana, Yonelinas, and Ranganath, 2007) have looked closely at the MT, attempting to classify different types of object representation and recognition. Further supporting evidence regarding the neural substrates of verb

concepts has been found using other imaging techniques like magnetoencephalography (MEG) (e.g., Pulvermüller, Shtyrov, and Ilmoniemi, 2005). A meta-analysis of imaging studies (Vigneau, Beaucoisin, Hervé, Duffau, *et al.*, 2006) confirmed the involvement of both the frontal and temporal lobes in semantic processing, although those authors made no apparent effort to isolate categories of words or to prove a double dissociation between nouns and verbs.

Indeed, the results from many imaging studies (including PET, fMRI, EEG, and MEG) have been rather inconclusive, with some finding little evidence for double-dissociated regions and others presenting conflicting evidence (as reviewed by Corina, Gibson, Martin, Poliakov, *et al.*, 2004). Corina *et al.* used cortical stimulation mapping on epileptic and tuberous sclerosis patients while showing them short films and asking them to name objects or actions. The results suggested selective impairments for action and object naming occur when different areas of the temporal lobe are stimulated. Interestingly, it was difficult to generalise across subjects, but differential disruption was found at an individual level for eleven out of thirteen subjects. For the three subjects with a very clear double dissociation, the site that was vulnerable during object naming was consistently anterior to the action-naming site. These results suggest that individual differences may play a significant role and account for some of the inconsistencies in the imaging results.

It is important to note that Corina *et al.* did not introduce transient lesions into more frontal areas and, thus, a role for the LPFC or PMC cannot be conclusively defined—or ruled out—based on their results; the use of films rather than photographs may also be confounding. Also, the use of surgical patients with no controls, although necessitated by their experimental procedure, means that the findings may not generalise well. Corina *et al.* did raise a number of methodological questions that could explain differences across other studies. In a language experiment, it is vital to consider whether the task is lexical, semantic, or even syntactic. Until the confusion across studies has been reconciled, the data do seem to indicate that the object and action concepts have separate semantic systems within the brain. How much the two systems overlap and exactly where each is located are not the cardinal questions for my model. The most consistent results—the lesion case studies and supporting imaging data—suggest a separation of noun and verb representations within the brain, which I will adopt.

For my model, the semantic systems for object and action concepts are the MT/IT and the LPFC/PMC, respectively. For the sake of simplicity, I will generally refer to

these regions as IT and LPFC. Damasio and Tranel characterised the identified regions as ‘mediation systems’, suggesting that they represent a “stage between the processing of a concept and the vocalization of word forms.” Thus, the complete storage of concepts and word forms involves interacting networks. According to this hypothesis, word forms are not explicitly stored in their entirety in these locations. Rather, these intermediary areas reconstruct a neural pattern that generates an explicit representation from sensorimotor information (Damasio *et al.*, 1996). Thus, the argument does not preclude a distributed representation of language within the brain, nor does it preclude the other regions highlighted by various imaging studies. It does, however, suggest that specialisation in discrete functional areas is divided on at least the level of conceptual representation.

The SWLM that I have used in my model is an appropriate representation of such a mediation system. As an artificial neural network, it does not store word information explicitly. It reconstructs words based on input patterns. The lesion studies cited above provide a rationalisation for using two separate SWLMs—one for objects and one for actions—with each modelling a different region within the brain. Such a separation begs the question of how the two areas are connected and how they interact, which is the idea that will be explored in the next section.

5.2 Neural Substrates of Intention Recognition

I will now discuss the brain systems for recognising actions and, in particular, recognising the intention of actions. This system is essential to Tomasello’s theory regarding the significance of intention recognition. I will argue that the prefrontal cortex (PFC) has a special role in representing intentions. In Section 5.2.1, I will discuss a model of PFC that describes how we overcome well-learned behaviour. In Section 5.2.2, I will talk about how intentions are represented in the PFC as persistent plans. I will then discuss the mirror system, a hypothesised mechanism for action recognition, in Section 5.2.3. Finally, Section 5.2.4 will discuss intention recognition and reiterate the relevance of such a system to the process of language learning.

5.2.1 Background: a model of PFC function

The primate frontal cortex is complex, occurring at its largest absolute size in the human brain (Semendeferi, Lu, Schenker, and Damasio, 2002). The subregion known as the PFC is the anterior area (see Figure 5.1.) It consists of three main areas: the

orbital PFC (OFC), the anterior cingulate cortex (ACC), and the dorsolateral PFC (DLPFC). While the first two subregions are found in the brains of many species, the DLPFC appears to be unique to primates (Striedter, 2005, p. 309). Various regions of the PFC are associated with many higher-order cognitive functions including decision-making (Damasio, 1995), memory (Dubois, Levy, Verin, Teixeira, *et al.*, 1995), attention (Dunbar and Sussman, 1995), procedural learning (Pascual-Leone, Grafman, and Hallett, 1995), planning (Sirigu, Zalla, Pillon, Grafman, *et al.*, 1995), response selection (Striedter, 2005, pp. 333–335), and inhibitory control (Striedter, 2005, p. 334). The PFC is well-suited to these tasks as it sends and receives information from many cortical sensory systems, motor systems, and subcortical structures within the brain (Miller and Cohen, 2001). Many researchers (e.g., Miller, 1999; Striedter, 2005) propose that the principal use of the PFC is to guide complex or unconventional behaviour.

Consider a simple model of stimulus-response (SR) behaviour, wherein one or more responses are possible for a given stimulus. Such a network is illustrated in Figure 5.2. The example that Miller and Cohen (2001) use to explain the model is one of looking both ways before crossing the street. Suppose a native resident is crossing the street (S1) in Canada (S2) where, as in many countries of the world, automobile traffic travels on the right-hand side of the road. The correct response is to look to the left (R1) before beginning to cross. After years of practice, this response becomes very well-learned and can be executed without conscious thought. However, if the same Canadian is visiting New Zealand (S3), where vehicles drive on the left-side, the correct behaviour when crossing the street (S1) is now to look right first (R2). Such switching very definitely requires conscious thought on the part of the pedestrian, but it cannot be modelled by associative SR alone; the well-learned S1-R1 mapping would be too strong.

Miller and Cohen proposed a **guided activation theory (GAT)** to explain how top-down influence can overcome well-learned behaviour. ‘Top-down’ refers here to the desires, goals, and intentions that guide behaviour—what is commonly referred to as conscious thought. According to their theory, the PFC is the tool of such cognitive control. The PFC stores active patterns of current goal states, or intentions. These patterns act as biasing signals that influence SR pathways to the mappings that achieve the intended goals. A SR diagram including PFC units appears in Figure 5.3. The PFC units are connected to the hidden units, which represent high-level association and premotor cortices, rather than primary sensory or motor ones (Miller and Cohen, 2001). In the earlier example, the vacationing Canadian in New Zealand would use

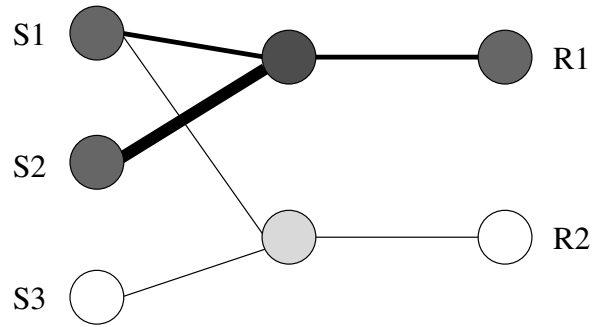


Figure 5.2: A diagram of simple stimulus-response behaviour (based on Miller and Cohen, 2001). Each stimulus (S1, S2, S3) can be followed by one or more voluntary responses (R1 and R2). Thicker lines indicate stronger connections. Intermediate, hidden units represent processing pathways between the two.

his PFC to bias the pathways to the response (R2) that is congruent with his goal of crossing the street (S1) while avoiding being hit by a car. The top-down signals from the PFC excite the weaker mapping until it wins the competition with the habitual tendency (R1). In short, the PFC is used to respond to stimuli in a task-relevant manner, potentially overriding habitual tendencies in order to choose a response that better achieves the current cognitive goal.

Evidence for GAT comes from such common psychological task paradigms as the Stroop task and the Wisconsin Card Sorting Task (WCST) (see Dunbar and Sussman, 1995 for more detail on these tasks). The classic Stroop test presents a subject with a word (e.g., *green*) written in an incongruent colour of ink (e.g., red). Subjects are required to either name the colour (considered a voluntary process) or read the word (considered an automatic process.) The WCST requires the subject to sort a pack of cards along one of three dimensions (shape, colour, and number.) The sorting rule will change periodically to see if the subject demonstrates perseverative errors when switching between tasks. Frontally-damaged patients have been shown to be impaired at both Stroop tests and the WCST, suggesting that they have difficulty overcoming well-learned responses (see Miller and Cohen, 2001 for a summary).

More recent studies of action sequencing (e.g., Zanini, Rumiati, and Shallice, 2002; Zanini, 2008) also support the GAT. The theory has not just been applied to explicit motor sequences, but also to other forms of context-guided behaviour such as language-related studies (Kerns, Cohen, Stenger, and Carter, 2004). Thus, experimen-

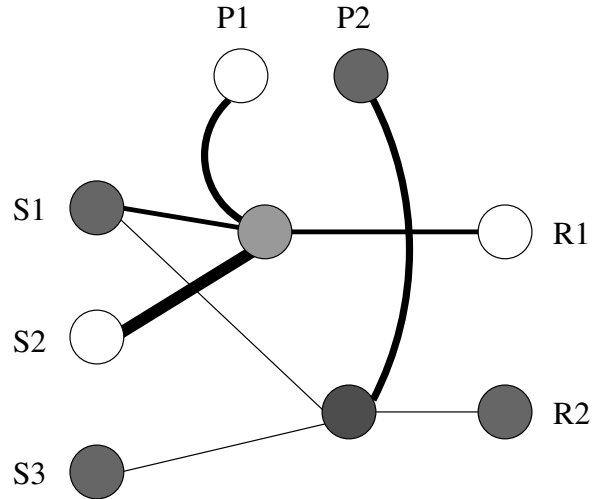


Figure 5.3: An illustration of the guided activation theory (based on Miller and Cohen, 2001). PFC units (P1, P2) excite hidden units to bias the selection of a desired response.

tal evidence supports the role of the PFC as a top-down bias in a variety of situations that demonstrate flexible and controlled behaviour. It is not surprising that humans, with a disproportionately large PFC—especially the DLPFC—compared to other primates, demonstrate the greatest flexibility for complex behaviour and the least degree of impulsivity (as reviewed by Striedter, 2005, pp. 334–335).

5.2.2 Intention Representation in the PFC

In order to be a viable substrate for goal planning, the PFC must be able to hold persistent representations and to maintain them when confronted with distractors. Early work with rhesus monkeys in a delayed response task showed PFC cells would begin firing at cue presentation and remain at an excited level throughout the delay, suggesting that the activity represented a focusing of attention and the maintenance of temporary memory storage (Fuster and Alexander, 1971).

A delayed matching to sample (DMS) experiment with rhesus monkeys analysed PFC and IT activity using cell recording (Miller, Erickson, and Desimone, 1996). The monkeys were presented with a sample stimulus and then one to four test stimuli in sequence. The monkey was rewarded for successfully identifying a match and ignoring the intervening distractors. PFC activity was sustained throughout the presentation of the non-match stimuli, whereas IT activity was not. IT activity displayed a more

stimulus-selective behaviour than that of the PFC, further underscoring the role of IT in object coding. Moreover, the evidence proves the ability of the PFC to maintain a representation in the face of distraction.

Averbeck, Chafee, Crowe, and Georgopoulos (2002) tested rhesus macaques while they drew sequential line segments to produce geometric shapes. Single-cell recording demonstrated that, at the beginning of the sequence, the different discrete movements were individually represented in the PFC. During error trials, the relative strengths more closely reflected the actual behaviour than the expected, suggesting that the PFC activation represents more than just an intended sequence.

More recent studies by Averbeck, Sohn, and Lee (2006) provided evidence of the PFC's role in planning, as well as further underscoring its involvement in dynamic switching of action sequences. The monkeys were trained to execute eight different sequences of eye saccades. After training, they were prompted to perform the sequences in an unpredictable order. A Gaussian decoding analysis was used to estimate which sequence was being represented in the PFC at any time, based on measured neuron activity. Because the monkey could not predict the sequence changes, there was an inherent trial-and-error aspect to its behaviour as it searched for the correct alternative in each new block. During this phase of uncertainty, the sequence represented by the activity-based probability calculations was found to change gradually from the previously correct sequence (i.e., from the previous block), to the currently correct one (i.e., for the new block.) The PFC activity for each sequence was still distinguishable in spite of the fact that different sequences shared common individual movements (e.g., moving eyes horizontally to the right was the first movement of four different sequences.) As the authors pointed out, "neural activity in the [PFC] related to a sequence of movements reflects not only the movement being made but also the sequence in which this movement is embedded" (Averbeck *et al.*, 2006). I will return to this interesting result later in this chapter.

Another study (Averbeck and Lee, 2007) applied a similar technique, but focused on error trials by studying neural activity during intertrial intervals. Consistent with previous results, DLPFC activity represented the sequence of actions to be taken, before the movements had even begun. Moreover, the error trials showed that the activity reflected the *actual* action plan that was executed rather than the one that would have been correct for the trial. As the authors point out, it is difficult to know whether the representation reflects a memory process for a previously-executed sequence, or a planning process for a future sequence. What is important, however, is

that the representation is maintained for longer than it takes to perform its constituent movements. Together with the Averbeck *et al.* (2002) study, these results suggest strong evidence for a correlation between persistent activity in the monkey PFC and plans for sequences of movements.

An fMRI study provided evidence of persistent neural activity in the human PFC (Srimal and Curtis, 2008). Subjects were scanned during spatial item recognition tasks with variable-length delays. Activation was found in frontal and parietal areas at cue onset and it was maintained for the entire duration of memory delays ranging from 7.5 to 13.5 seconds, well within the temporal resolution of a blood oxygenation level dependent fMRI (approximately 1 second.) Because preparation cues were presented for 1.5 seconds before trial stimulus, it can be assumed that the measured activity was more than simply an index of focused attention.

Thus, the PFC appears to hold a persistent representation of task execution that exceeds simple attention, resists distraction, and involves an (at least temporary) form of information storage.

I have now discussed, as some length, the role of the PFC in overcoming well-learned behaviour. However, I have yet to address what purpose it might serve to learning in the first place. For instances of complex behaviour, the outcome of a sequence of actions can be delayed from the individual steps by some time. Miller and Cohen also discuss how the active maintenance of information in the PFC over time could be one tool for learning associations between actions and delayed outcomes or consequences. Dopamine (DA) neurons in the basal ganglia have been shown to fire spontaneously upon the delivery of an unpredicted, rewarding stimulus (Schultz, 1998). Over time, the DA neurons begin to anticipate the reward and will fire during events that have led to reward in the past. They begin to code a prediction signal, with activity attenuated based on whether or not the anticipated reward is realised (see Miller and Cohen, 2001 for a review.) This behaviour is a neural example of the sorts of reinforcement learning mentioned in Chapter 3 (e.g., Triesch *et al.*, 2006). An active representation over time is necessary to realise this form of learning, and there is evidence that the PFC conveys information about expected rewards (Miller and Cohen, 2001). The release of DA, in turn, is hypothesised to serve as a gating signal for the update of goal representations in the PFC (Miller and Cohen, 2001).

5.2.3 The Mirror System: action recognition

Closely related to the system of complex behaviour is an action recognition system. When we make a cup of tea, one step in our plan is to reach for the kettle that we will use to boil the water. When we see another person reach for a kettle, we can easily infer that he/she intends to boil water with the likely intent of making a cup of tea. Anticipatory eye movements are an important component of this system. Flanagan and Johansson (2003) studied eye movements during performed and observed goal-directed actions. When subjects stacked small blocks, their eyes would look to a block before their hand reached it to pick it up and also to a new location before their hand had placed it there; a similar pattern of eye movements was demonstrated when the subject watched an experimenter perform the actions. Thus, when I reach for the kettle, my eyes will look at it before my hand reaches it. When I watch you reach for the kettle, my eyes will anticipate your goal and look to the kettle before you have completed the motion. Falck-Ytter, Gredebäck, and von Hofsten (2006) found that such proactive goal-directed eye movements were not apparent in six-month-olds, but are demonstrable from at least 12 months of age.

The field of action recognition research was revitalised by the discovery of **mirror neurons** in the premotor cortex of the monkey brain (Gallese, Fadiga, Fogassi, and Rizzolatti, 1996). The mirror neuron system (MNS) theory has since been expanded to include a network of discrete brain regions. One particular area, F5, is a premotor area with a somatotopic organisation that includes representations of the mouth and hand. Using single-cell recording, such neurons have been found to react very specifically to manual actions like grasping or reaching (Rizzolatti, Fadiga, Gallese, and Fogassi, 1996). Some neurons were found to not only discharge when performing an action, but also when observing another animal performing that same action—hence the term ‘mirror neuron’ (di Pellegrino, Fadiga, Fogassi, *et al.*, 1992; Gallese *et al.*, 1996).

Further evidence suggests that the MNS surpasses simple action recognition to encompass some understanding of intentions. Fogassi, Ferrari, Gesierich, Rozzi, *et al.* (2005) recorded from neurons in the monkey rostral inferior parietal lobule (IPL) during grasp actions. Food (or another small object) was grasped and either brought to the mouth or placed in a container. The majority of the neurons coded the final goal, firing differentially depending on which of the two actions was performed. Consistent with their classification as mirror neurons, the firing rates were similar regardless of whether the monkey executed the grasp or observed an experimenter performing the same action. The authors controlled for a number of factors such as reward, arm kinematics,

applied force, object grasped, and even the final location. The main factor affecting the behaviour of the neurons was the final motor act, or the intention behind the sequence of movements. Interesting parallels may be drawn with the Averbeck *et al.* (2006) study mentioned previously, wherein PFC activity related not to just individual movements, but also to the sequence in which those movements were embedded. Because each sequence began with the same action, grasping the food, the MNS is hypothesised to have a predictive component that constructs an internal representation of a likely action sequence before it has been completed. Such predictions are a necessary component of an intention recognition system. Seeing an action performed ignites a representation of the perceived intention.

Single-cell recording is obviously unethical and problematic with human subjects. However, researchers have been very curious to discover whether or not humans have a homologous system. Studies investigating the possibility of a human MNS have been conducted using a variety of different imaging techniques as indirect measures (e.g., Fadiga, Fogassi, Pavesi, and Rizzolatti, 1995; Rizzolatti, Fadiga, Matelli, Bettinardi, *et al.*, 1996; Binkofski, Buccino, Posse, Seitz, *et al.*, 2000; Pfurtscheller, Neuper, and Krausz, 2000). Such a system is of particular interest because F5 is thought to be a homologue of Broca's area in the human brain, a region closely associated both with hand movements and with speech (Gallese *et al.*, 1996); the region appears to have no *non-primate* equivalent (Striedter, 2005, p. 307). A purported human MNS is also supported by earlier theories, such as the motor theory of speech perception (Liberman and Mattlingly, 1985).

The sheer quantity of imaging studies definitely supports the hypothesis of a human MNS. Using MEG, Hari, Forss, Avikainen, Kirveskari, *et al.* (1998) demonstrated that activity in the precentral motor cortex will rebound in a similar manner, but to a lesser degree, after observation of object manipulation as it would when performing that same action. Muthukumaraswamy, Johnson, and McNair (2004) used EEG to show that a brain rhythm, called mu, that modulates during a grasp action will also modulate during observation of the same action by another human. When subjects observed or imitated meaningless finger movements, fMRI revealed that common brain regions—in Broca's area and in the right anterior parietal cortex—would be active (Iacoboni, Woods, Brass, Bekkering, *et al.*, 1999). These data not only suggest the existence of a homologous human system, but also that it may have slightly different properties because the monkey MNS only seems to respond to goal-directed actions.

Based on my discussion of the MNS, I propose that a representation in the brain

when there is an intention to act is very similar to the one when that same intention is recognised in others. This proposal raises questions of agency. If these representations are so similar, how does the human (or monkey) know the difference between when he intends to grasp an object himself and when he recognises another intending to grasp the object? I assert that the distinction suggests a switch between ‘modes’ of action execution and action recognition.

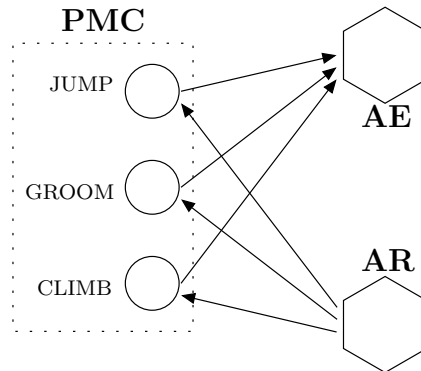


Figure 5.4: A simple model of an integrated action execution (AE) and action recognition (AR) system. Each node in the PMC represents a possible action.

I will now discuss a very basic model of how an integrated action-execution/action-recognition system could work by switching between different modes. An example is shown in Figure 5.4 for a monkey with a very limited repertoire of possible actions: jumping, grooming, and climbing. Each of these actions is represented by a node in the PMC layer. Each action node sends output to an action execution (AE) system and receives input from an action recognition (AR) system.

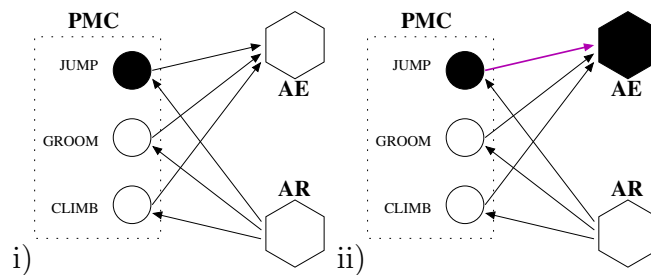


Figure 5.5: An example of action execution: the intention to jump.

Figure 5.5 shows an example of action execution, wherein the monkey forms an intention to jump. At step *i*, the jump plan is represented, with the jump node activated

in the PMC. At step *ii*, activation feeds forward to the AE system and the action is executed by the motor system—in other words, the monkey jumps.

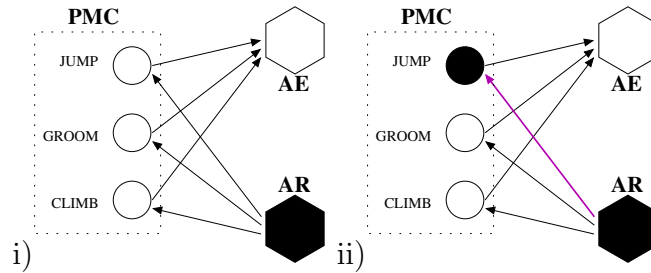


Figure 5.6: An example of action recognition: recognising that a conspecific intends to jump.

Now consider the action recognition example shown in Figure 5.6. The monkey is observes a conspecific, activating the AR system at *i*. At step *ii* the observed monkey's intention to jump is recognised by the first monkey's MNS; activity is fed back to the PMC and the monkey forms a representation of JUMP just by watching another perform the action. This very basic example illustrates how the action-execution and action-recognition systems could overlap a great deal in their underlying neural circuitry, but still remain distinct. Now that I have argued for such a system, it is essential that I fit it back into the context of communicative intentions.

5.2.4 Intention Recognition and Neurolinguistics

Knott (2008) has sketched a neurolinguistic account of the planning and intention recognition system to explain how humans represent an entire proposition within the brain. These propositions are described by entire sentences consisting of a subject, verb, and object; e.g., “Mommy grabs the ball.” Observing the proposition or event requires the observer to proceed through a stereotypical order of sensorimotor processes: first attending to the mother, then attending to the ball, and finally attending to the grasping action. This is supported by eye-gaze studies (Webb, Knott, and MacAskill, 2008), which find a characteristic sequence of agent-target saccades in subjects observing a video of reach-to-grasp actions. Because the order (subject-object-verb) of observation is not directly related to the spoken order (subject-verb-object), the correlation is assumed to be more than an artifact of English. It is also important to note that isolated, unordered representations (i.e., MOTHER AND BALL AND GRASP) are not sufficient for binding reasons. That is, an accurate observation necessitates a representation of

who did what to whom, as illustrated by the more general case “the boy kissed the girl.”

Knott proposes that the proposition is represented internally according to the same characteristic sequence demonstrated by the eye movements: the agent of the action, then the patient, and finally the action itself. As such, a mapping is proposed between the observed concepts and their internally-represented thematic roles. The ordered concepts representing the entire proposition can be thought of as an intention, held active by WM in the PFC.

Consistent with the existence of an hypothesised human MNS, an observer who recognises such an intention creates an internal representation similar to the one that he/she would use to execute the same act. If a child watches his mother chase after a rolling ball, the sequence would include the mother, then the ball, and finally the motor scheme associated with TO CHASE. The individual elements of the sequence roughly correspond to what Langacker (1987) termed ‘cognitive events’.

Once the proposition is complete, it can be neurally rehearsed, consistent with the conceptualisation of a “plan” as a simulated potential act (Gallese and Lakoff, 2005). The child could then use it to execute his own act of running after a ball. Of course, many levels of abstraction are possible. Thus, a child might watch a sea lion chase a horse.¹ Anatomical differences would prevent him from precise imitation, but he could use his own mirror system to represent the action witnessed at a higher level. Even if the sequence is not actually imitated, of course, the neural rehearsal may be used as a form of mental replay to recall what was witnessed. The theory proposes that, during those internal rehearsals, a series of transient neural activations occurs. First there is a representation of the agent (i.e., the mother) in IT, followed by an IT representation of the patient (the ball), then an action representation (chasing or running) in the PMC. The sequence plan is assumed to be tonically active throughout the entire series of transient activations. Each relevant concept, then, is represented twice: transiently in the IT or PMC and tonically in the PFC.

The idea that an entire proposition is represented in PFC must be considered from the perspective of the dissociations found in the semantic retrieval studies (e.g., Damasio and Tranel, 1993). The role of the LPFC as a semantic mediation system for actions suggests direct access to the verb lexicon. Thus, rich relationships between PFC action

¹While such an occurrence may seem improbable, I can personally attest that it is not impossible. I witnessed the rather unlikely event, myself, while working on this thesis. A child is likely to witness much more improbable acts on television and endeavour to imitate some of them.

representations and verbs are proposed, but only weak ones between PFC object representations and nouns. Knott proposes that the latter connections contribute to the verb morphology, consistent with imaging studies that suggest morphological processing involves frontal-temporal interaction via the dorsal route of the arcuate fasciculus (Tyler and Marslen-Wilson, 2008) . In order to correctly conjugate the verb (e.g., “the sea lion *chases* the horse”), the speaker needs to have a representation of at least some aspect of the agent. Finally, consistent with the Damasio and Tranel’s notion of a mediation system for object semantics, there must be rich connections from IT object representations to the noun lexicon.

Knott’s hypotheses fit well with the theories of Gallese and Lakoff (2005). They propose that the sensorimotor system used to execute actions is the same region of the brain that characterises sensorimotor concepts, as well as more abstract ones. For example, the concept for GRASP is represented by the same sensorimotor system that is used to perform a grasp. To argue their case, they review evidence of ‘action-only’ neurons—those that fire only when a purpose-directed grasp is actually performed—and ‘canonical neurons’—those that fire when a grasping action is carried out, as well as when an object is observed that *could* be grasped—in area F5 of the monkey brain. (Canonical neurons behave differently from mirror neurons, which will not fire at the mere sight of an object that can be manipulated.) If the monkey grasps a small object, certain action-only neurons will fire along with certain canonical ones. If the monkey observes a similar small object, the same canonical neurons will fire. If the monkey observes an object that would require a different form of grasp (e.g., something much larger), different canonical neurons will fire. To the authors, this evidence suggests that the sight of the graspable object automatically triggers the concept for GRASP—essentially a simulation of a plan to grasp the object. A similar proposal explains the behaviour of mirror neurons. Performing an act and observing that act being performed (a simulation) elicit common neural activity because there is overlap in the neural substrates for action and conceptualisation. Just as with mirror neurons, neuroimaging has found indirect evidence of action-only and canonical neurons in the human brain (see Gallese and Lakoff, 2005 for a review.)

Gallese and Lakoff propose that the sensorimotor system conceptualises more than just concrete actions like grasping, but also more abstract ones. Further, the primary use of language is the expression of concepts and, logically, it would use the same system. An equivalent conceptual system could exist in other location of the brain and be dedicated to expression of concepts, but such duplication would be contrary to a par-

simonious view of the brain. Instead, these theories propose that the human brain has a rather tightly coupled system for performing, observing, imagining or remembering, conceptualising, and expressing actions and other concepts.

5.3 A Schematic Model of Action Execution and Recognition

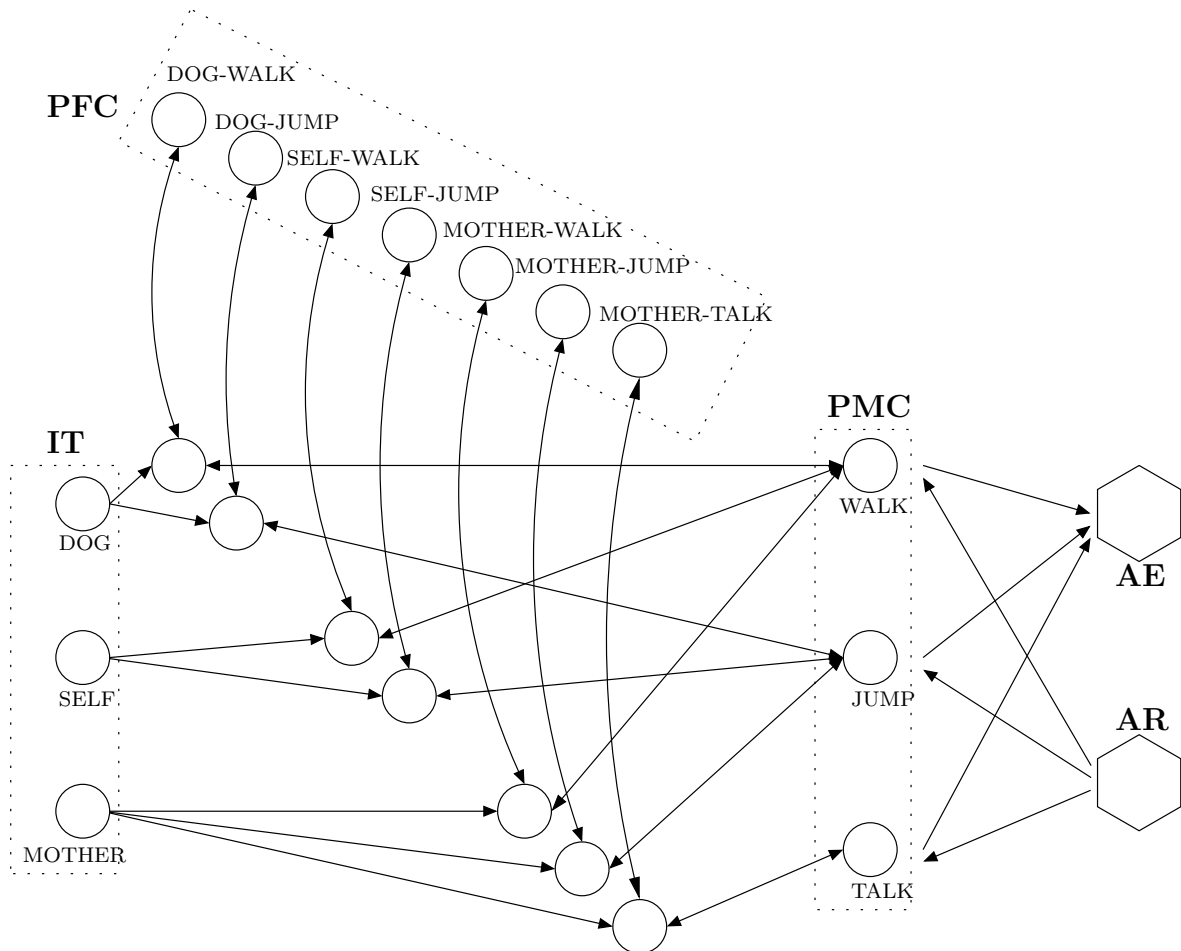


Figure 5.7: An adaptation of Miller and Cohen’s GAT, combined with the AE/AR system. The PFC, IT, and PMC layers are connected by hidden layer. Inhibitory connections between hidden neurons are not shown.

I will now discuss how the above-mentioned research can be used to improve upon my own model. The first building block is a simple model of action recognition and

execution, of the sort that could identify actions such as DOG-JUMP or MOTHER-TALK. This model takes the form of an adaptation of Miller and Cohen’s GAT, as shown in Figure 5.7. Entire propositions are represented in the PFC layer, which performs the sort of top-down biasing proposed by Miller and Cohen. Objects are represented in the IT layer. Actions are represented in the PMC layer, which is connected to systems for action execution and action recognition. A hidden layer connects the three main layers (PFC, IT, and PMC.) Nodes are labelled with sample action and object concepts for the sake of illustration but, of course, even an infant brain would be expected to be able to form a much greater variety of representations. In a naïve system, each agent could theoretically perform every possible action. In practise, however, certain agents will never engage in certain behaviours—DOG-TALK, for example. The connections between these sensorimotor pairs will weaken over time until they are effectively pruned. Figure 5.7 presumes prior experience and such connections are not even shown.

As conceived by Miller and Cohen, the GAT is primarily a model of action execution. The idea is that the PFC can hold planned actions, triggered by particular perceptual stimuli. Consider the case shown in Figure 5.8, that of me forming my own intention to jump. In this simple case, the triggering stimulus for the agent’s own action is the activation of SELF—which we can understand as some sort of general decision to act, i.e., to be an agent. The PFC unit SELF-JUMP biases the pathway from SELF to JUMP. In step *i*, the IT unit for SELF is activated, as is the PFC unit for JUMP. At *ii*, the activation from the IT unit feeds forward and activates two units in the hidden layer. However, the PFC unit is also connected to one of those units and its contribution adds to that of the IT input, causing one of the hidden units to be more strongly activated than the other one. Because each hidden unit has lateral inhibitory connections to every other unit in the hidden layer (not shown in the figure)², this strongest unit will inhibit the less activated unit and ‘win’ to become the only activated unit in the layer. At step *iv*, the winning hidden unit feeds its activation forward to the PMC, which sends it to the AE system in the motor cortex.

Next, I will describe a case of action recognition. As discussed at some length earlier in this chapter, humans have tightly-coupled action and intention recognition systems. When I see a dog crouch by a fence in a certain way, I can anticipate that it plans to jump the fence before the action is complete. Moreover, because of my hypothesised MNS system, observing the dog jumping will cause activation in my own brain that overlaps with the activation that would be present if I were to perform (or even just

²See Appendix A for more on lateral inhibition.

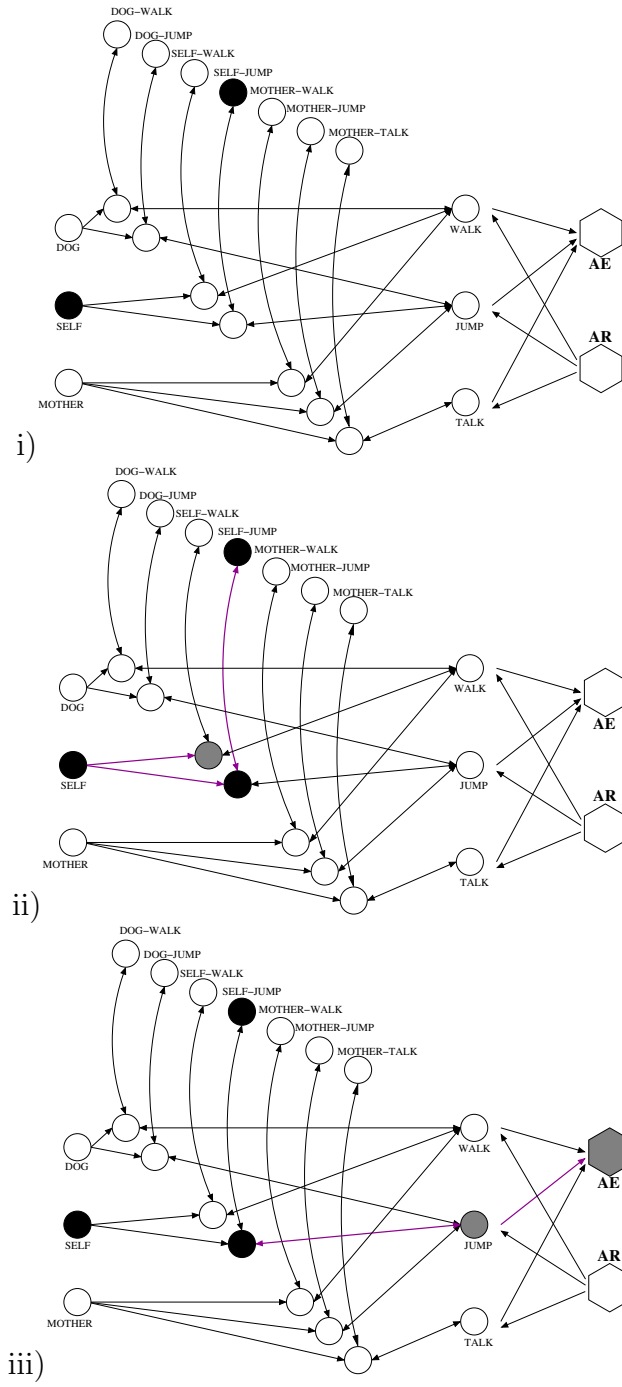


Figure 5.8: Example behaviour of the system for a representation of SELF-JUMP.

plan) my own jump. In other words, observing a jump creates a representation in my mind that is very similar to the representation I would have if I were jumping. This representation allows me to abduce what an intended act is, often before it has been completed. An infant in the second year of life is capable of jumping and has at least begun to develop the ability to anticipate actions and recognise intentions, so I assume that these abstractions also hold true for them.

To support this theory of action recognition, I assume that the connections between the PMC and hidden layer are bidirectional, as are those between the PFC layer and the hidden layer. Another example, shown in Figure 5.9, demonstrates how this model performs action recognition. The IT unit for DOG and the PMC unit for JUMP are activated at step *i*. At *ii*, activation in the IT layer feeds forward (along highlighted connections) to two units in the hidden layer. Activation from the PMC layer feeds back to three units in the hidden layer. One hidden unit is activated by both the IT layer and the PMC layer, making it the strongest and causing it to win the competition with the other hidden units. After a winner is declared, at *iii*, activation is fed up to the PFC layer. Thus, the entire proposition DOG-JUMP will now be represented in the brain of the observer. Importantly, the observer is representing the dog's *intention* to jump, as well as the fact of its jumping.

5.4 An Intention Recognition Model of Word Learning

Now that I have described how the model can work for action execution and recognition, I will return to my original focus. I am more interested in how intention recognition can facilitate word learning. Intention recognition can be thought of as a generalised case of action recognition. Figure 5.10 shows the result of adding a JAF and two SWLMs to the model from Figure 5.7. As before (see Figure 4.2), a gating layer is responsible for activating or inhibiting the SWLMs; inhibitory connections are not shown. The word learning gate is controlled by the PFC layer. The thick, bold connections represent information-sharing that carries conceptual sensorimotor information to the inputs of the SWLMs, when they are active. Note that the gating layer does not differentiate between nouns and verbs. It either activates both SWLMs or inhibits both of them. The noun network receives object concepts via the connection from the IT layer and the verb network receives action concepts as input from the PFC layer.

To summarise, the PFC layer will hold representations of propositions that are

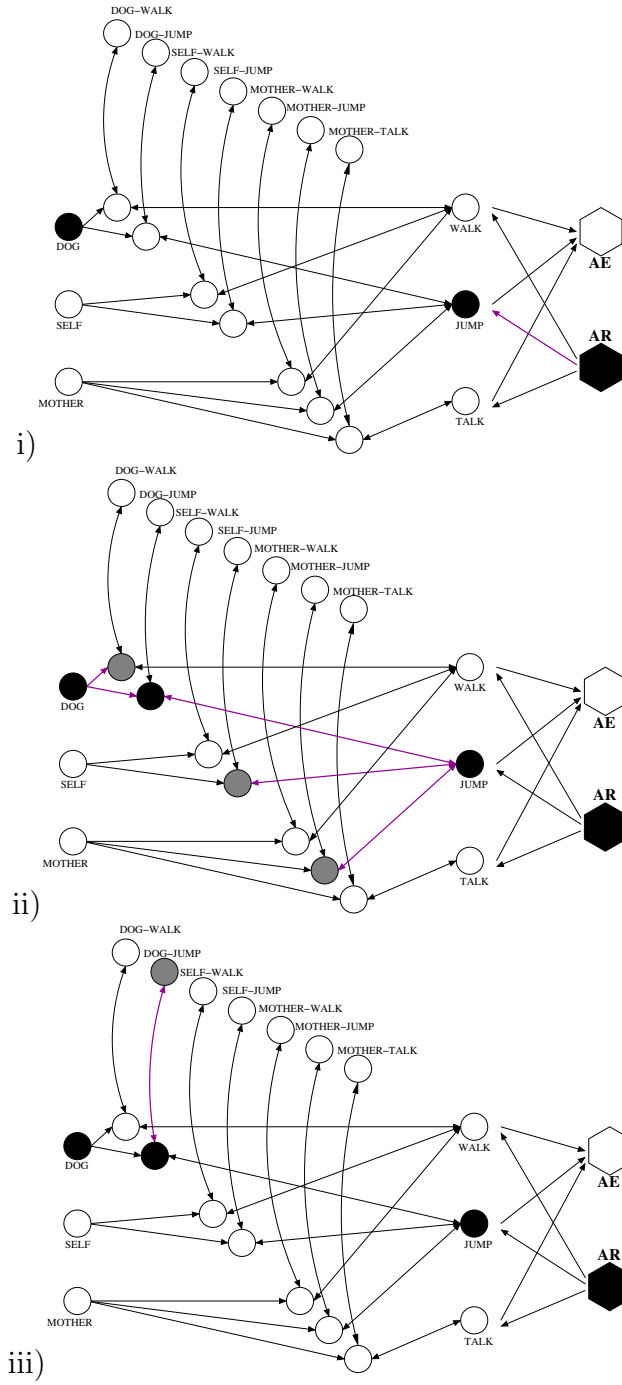


Figure 5.9: Example behaviour of the system for recognising that a DOG is performing the action JUMP.

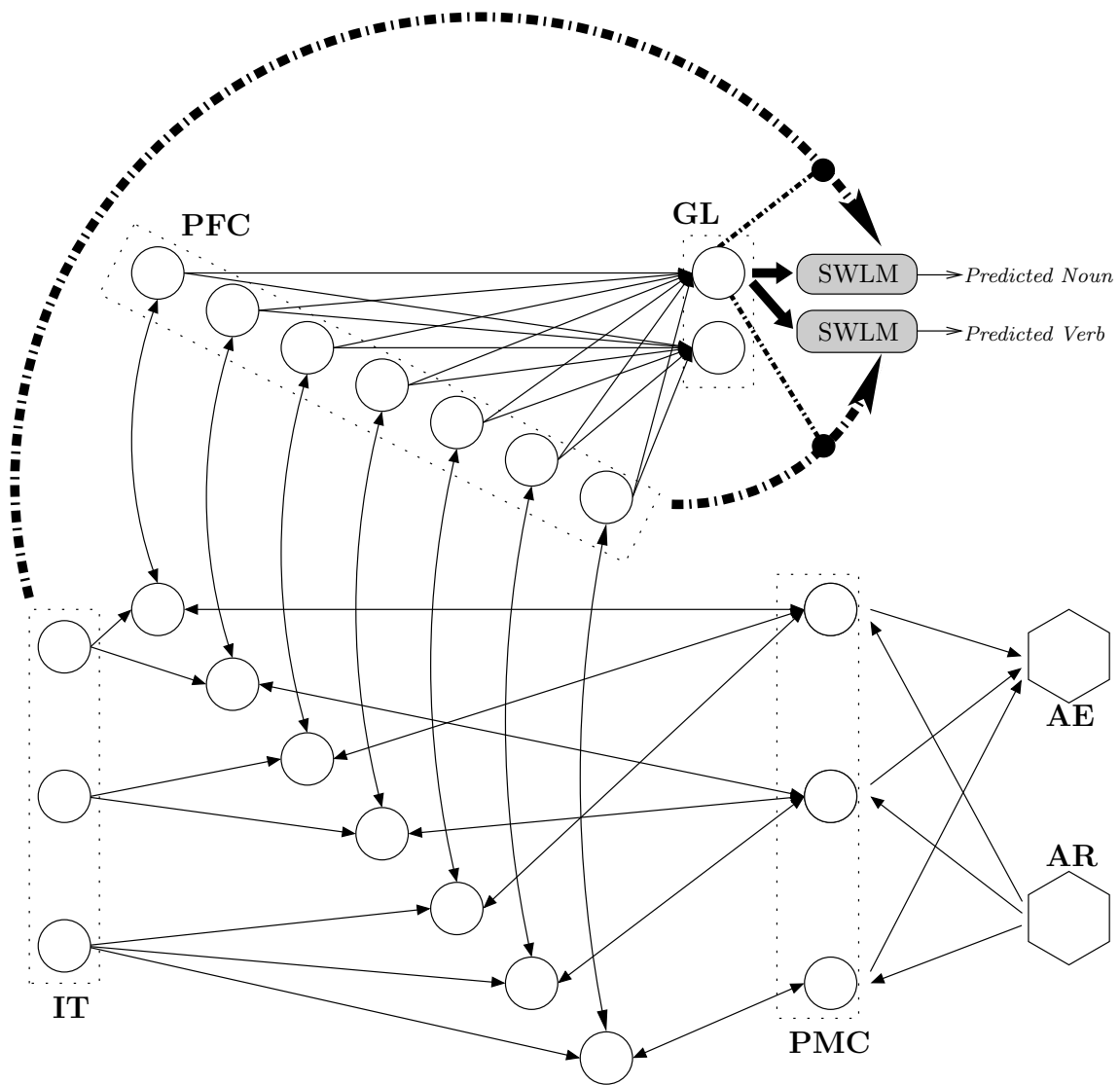


Figure 5.10: Illustration of an intention recognition word learning network with a joint attention filter and two simple word learning modules.

witnessed by the infant. Recognising a communicative situation will consist of the activation of MOTHER-TALK in the PFC layer, based on activity propagated from other parts of the network. This extra step was implicit in the model of Chapter 4. Responding to the joint attention to process the word-learning situation will replace³ the PFC representation of MOTHER-TALK with a representation of the content of the communicative intention (e.g., DOG-JUMP), in order for that content to be active in the network at the time the object-action pair is passed from the IT/PFC layers onto the SWLMs via the dotted-line connections. As before, a delay is assumed in the transmission of this information (i.e., MOTHER-TALK is recognised and the object-action pair of the *next* time step, DOG-JUMP is passed to the SWLMs.)

5.5 Experiment D: Learning Nouns and Verbs

Before I explore the more advanced model in any detail, I will describe simple simulations to examine the effect of adding a second SWLM for verb learning. Although many of the experiments that formed the basis of my initial model (e.g., Baldwin, 1991), were focused on the learning of nouns, many of the same principles apply to communicative situations intended to teach novel verbs (e.g., Akhtar *et al.*, 2001—see also Tomasello, 2003, pp. 71–78 for a discussion of using linguistic context to learn a novel verb).

In the previous chapter, I simulated noun learning by giving the networks input dominated by object concepts. Adding a separate network that learns verbs based on input dominated with *action* concepts would be trivial. However, because I now plan to use a vocabulary that is evenly split between nouns and verbs, the input should logically also be evenly distributed. Such a distribution of nouns and verbs inherently increases the level of noise in the input to the individual networks.

The reason for the increased noise is that each training trio consists of a noun concept, a verb concept, and a lexical item. Even in the correlated cases, the word corresponds to the action *or* the agent but not both (e.g., DOG-JUMP-*jump* or DOG-JUMP-*dog*.) Now that the vocabulary is split evenly between nouns and verbs, there is an approximately 50% chance of the word being one or the other. Thus, there will be greater variability in the training input, even when the JAF is active. However, as the example in Table 5.1 demonstrates, there is still a clear signal for the network to recognise amidst the noise. As such, it should still be possible to map the correct

³A discussion of a mechanism by which the proposition might be replaced appears in Appendix A.

Training Trio	Mapping to Learn
DOG-JUMP- <i>dog</i>	DOG→ <i>dog</i>
DOG-WALK- <i>dog</i>	DOG→ <i>dog</i>
DOG-WALK- <i>walk</i>	DOG→ <i>walk</i>
DOG-DANCE- <i>dance</i>	DOG→ <i>dance</i>
DOG-SCRATCH- <i>scratch</i>	DOG→ <i>scratch</i>
DOG-RUN- <i>dog</i>	DOG→ <i>dog</i>
DOG-JUMP- <i>dog</i>	DOG→ <i>dog</i>

Table 5.1: Sample training input (first column) and the mappings that the noun SWLM would try to learn (second column.) Essentially, the network will be instructed that DOG maps to *dog* more often than it sees it mapping to any particular verb.

output over time (i.e., the noun network will see more examples of DOG-JUMP-*dog* than it will any other trio involving DOG-JUMP.)

5.5.1 Methods

A vocabulary of 15 nouns and 15 verbs is used. The number of hidden units is increased to 30 (i.e., $2 * 15$ words) in each case. The JAF filter was activated in each network after it could correctly map at least 2 words. Training continued until each network produced 100% correct mappings for its entire vocabulary. One training set was generated with a CDS of 0.1. Training was repeated ten times for each network, resulting in 20 result sets (2 parts of speech X 10 training runs.) There were no statistical-only (i.e., without the JAF) regimes run in this experiment.

Although some researchers suggest that children learn nouns more easily than other parts of speech in certain languages and verbs more easily in others (Tomasello, 2003, pp. 45–50), there is no consensus in the literature (see Woodward, 2000 for a review of some dissenting opinions). Therefore, my design makes no *a priori* assumptions about the relative difficulty between learning nouns and verbs.

5.5.2 Results

Data were collapsed across each set of ten training runs by taking the mean of the total vocabulary at any epoch, yielding one data set for nouns and one for verbs. Results

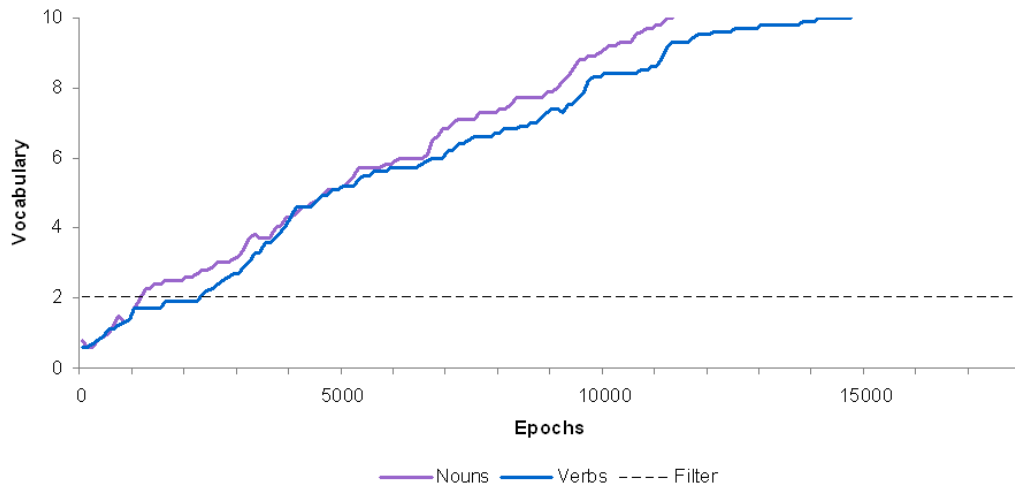


Figure 5.11: Learning nouns and verbs in Experiment D, presented as number of words learned over successive epochs.

are shown in Figure 5.11 for the first ten words learned. A single-factor ANOVA across all 15 words confirmed that there was no significant difference ($p > 0.75$) in learning rate between the two networks.

5.5.3 Discussion

As hypothesised, the additional noise from an increased proportion of verbs in the vocabulary did not impede either of the networks. Essentially, the noun network learns to treat the action inputs as noise, and the verb network learns to treat the object inputs as noise. Based on these results, I will now proceed to a more complicated network that is able to learn nouns and verbs using two SWLMs.

5.6 Experiment E: Learning When to Enable the Joint Attention Filter

5.6.1 Introduction

In all of my simulations so far, joint attention has been realised via a filter with binary states; it is magically switched on at a certain point in development. While this has been a useful simplification for focusing on other issues, it is important to also describe a method whereby the network can learn how to toggle the joint attention filter

itself. Like all cognitive skills, the ability to recognise the significance of communicative intentions and selecting to learn words in those situations would develop over time. For the next experiment, I will propose a method whereby my network *learns* when to gate the JAF that enables word learning.

5.6.2 Methods

In terms of my model (see Figure 5.10), learning when to activate the SWLMs amounts to modifying the weights of the connections between the PFC layer and the gating layer. Ideally, the network will converge on a weight space that would have MOTHER-TALK turn on the SWLMs most of the time and all other PFC nodes turn it *off* most of the time. Such a goal could be easily achieved using supervised learning; in real life, of course, there is no explicit training signal for an infant.

The reward-based learning scheme mentioned earlier (see Section 3.2.1), suggests an intriguing starting point. The network will be trained to predict reward. The reward in this case will be based on the infant receiving some form of positive feedback when a concept is correctly mapped to a word. This reward could be anything as simple as a smile or cheer of encouragement from the mother. However, a full-blown TD learning implementation (see Triesch *et al.*, 2006) is more complicated than the current problem requires because the sequences never involve more than two steps. Essentially, the network must learn that the MOTHER-TALK is a good predictor of these rewards, thus encouraging the SWLMs to be enabled in situations where there is a good chance of reward and disabled when there is not.

The network is trained using the same sorts of concept-concept-word trios (e.g., DOG-JUMP-*dog*) used to train previous networks. At each step, activity propagates forward from the PFC layer and a winning node is chosen in the gating layer. If the enabling node is the winner, the SWLMs are enabled for the next time step and the trio from that next time step will be fed to the appropriate SWLM. An example is shown in Table 5.2. Thus, at $t=2$, the SWLMs are enabled. At $t=3$, CAT is given as input to the noun SWLM and *cat* is the expected output.

Now, I will explain the reinforcement algorithm used to modify the connections from the PFC layer. Because of the single-step look-ahead, the connections at time $t-1$ are always adjusted based on the outcome at time t . There are three possible results, with the corresponding behaviours described below:

1. The SWLMs are not active at time t . Learning occurs according to classical

t	0	1	2	3
Noun Concept	DOG	DOG	MOTHER	CAT
Verb Concept	RUN	SLEEP	TALK	JUMP
Word	<i>walk</i>	<i>sleep</i>	<i>dog</i>	<i>cat</i>
Expected Result (SWLMs)	inhibited	inhibited	enabled	input: CAT- <i>cat</i>

Table 5.2: A sample timeline for learning when to enable the JAF.

Hebbian principles, with the connection between any two nodes strengthened in small proportion, $learn_{default}$, to their mutual activation at $t-1$.

2. An SWLM was active and it performed a correct mapping, resulting in a reward for the system. The Hebbian learning is attenuated by a strongly positive scale factor, $learn_{reward}$.
3. The SWLM produced an incorrect mapping (i.e., the word does not match the concept), generating a punishment. The Hebbian learning is attenuated by a negative proportion, $learn_{punish}$, weakening the connection from the PFC node that enabled the SWLM at $t-1$.

Connections from the PFC layer to the gating layer are initialised with weights of small random strength. As such, random exploration will tend to enable the SWLMs around 50% of the time at the beginning (there are only two possibilities.) Because the non-JA trios tend to be uncorrelated (e.g., KIWI-YAWN-*walk*), a punishment results if they activate the SWLMs, discouraging such behaviour in the future. In order to assess the validity of the reinforcement scheme, while minimising the influence of other possible factors, the SWLMs are pre-trained (as in Section 4.2.2) with the full vocabulary. This unrealistic concession will be removed in the next experiment, after my reinforcement algorithm has been validated.

In terms of analysis, each input trio can be classified by comparing its actual effect on the SWLMs to its expected behaviour. If MOTHER-TALK activates the SWLMs this is considered a *true positive*. If any other input inhibits the SWLMs, this is also desired behaviour and is classified as a *true negative*. Because there are only two choices—switch SWLMs on or off—there is no need to count the other cases. By categorising the results this way, it is possible for me to quantitatively assess the performance of the simulation. If the algorithm is working, the number of both true positives and true negatives will increase from epoch to epoch.

The vocabulary consists of 10 nouns and 10 verbs. A training set was generated using a CDS of 0.1 and CC of 0.0 (see 4.3.2 for more information.) The successive trios from the training set are presented and connection strengths are adjusted according to each reward signal. Such an implementation, known as ‘online learning’, involves an immediate update of the weights after each training example. An epoch is defined as one pass through the entire training set. At the end of each epoch, an error term is computed. The error term (ϵ) is expressed in Equation 5.1, where p is the number of punishments, f_{neg} is the number of false negatives, and N is the total number of input patterns. Once the error term drops below a certain threshold (0.1), training stops. Training was repeated with 10 different networks on the same training set.

$$\epsilon = \frac{p + \frac{f_{neg}}{2}}{N} \quad (5.1)$$

It is important to note that the number of true positives is not included in the above calculation. As such, it is possible for a network to ‘converge’ even if MOTHER-TALK is *never* activating the SWLMs. However, the network should never reach such a state if the algorithm is working. Therefore, it is an effective exit strategy.

Network Parameter	Value
$learn_{default}$	0.04
$learn_{reward}$	0.2
$learn_{punish}$	-0.04

Table 5.3: Network parameters used in Experiment E.

It is difficult to speculate about what would constitute ‘realistic’ values for network parameters, such as the reward/punishment proportions. Through trial-and-error, I eventually arrived at the arbitrary values shown in Table 5.3.

5.6.3 Results

For this experiment, it is sufficient to show that the algorithm learns appropriately. As such, the number of true positives and true negatives is expected to show an upward trend. It is important to realise that, unlike many gradient descent algorithms in neural networks, there is no guarantee of monotonically increasing numbers. The reason for this behaviour is the random exploration built into the architecture. For example, say KIWI-YAWN inhibits the SWLMs at epoch 0. No punishment results, so connection

strengths undergo only minor modifications. At epoch 1, KIWI-YAWN could easily activate the SWLMs because there is still no strong bias for either behaviour. In this case, one less true negative would be counted at epoch 1 than occurred during epoch 0.

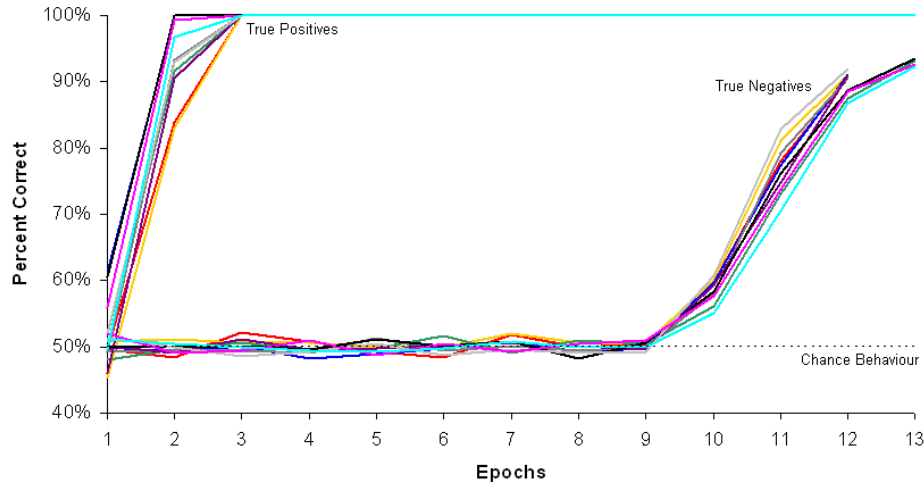


Figure 5.12: Experiment E. Learning when to filter, based on reinforcement from pre-trained word learning networks. A training example is considered ‘correct’ when it produces the expected behaviour in the network.

Results for the ten training runs are shown in Figure 5.12, including a dotted line to represent chance behaviour. The percentage of true negatives begins around chance level and then shows a clear upward trend as the network learns to ignore noise, reaching 90% accuracy in 12 or 13 epochs. The percentage of true positives is optimised by the end of the third epoch, suggesting that the reliable indicator of reward (MOTHER-TALK) is recognised quickly. This steep climb is likely a side effect of the online learning approach. Because the SWLMs are perfectly trained, rewards come quickly and accurately to the system.

5.6.4 Discussion

Based on the data, the algorithm appears to succeed with very little variability in performance. Again, a rigorous examination of the results is not warranted. The simulation was useful for demonstrating the utility of the reinforcement learning scheme, but it is merely a stepping stone to more interesting work. So far, I have investigated

the case where the filter is perfectly attuned before word acquisition begins (Chapter 4) and also the case where the filter emerges after a small vocabulary (see above.) However, neither of these extremes is very plausible. More accurately, the two skills would develop in parallel. I will now document such an approach.

5.7 Experiment F: Bootstrapping the Joint Attention Filter

5.7.1 Introduction

As I have mentioned numerous times, children do not jump from zero ability to a perfect command of language in one step. Like so many other cognitive skills, language *develops*. There is also a developmental trajectory for the recognition of communicative intentions. In previous experiments I have demonstrated the following:

1. The ability to use joint attention to identify communicative situations will speed up word learning by filtering out irrelevant ambient speech (see 4.3.3).
2. Joint attention filtering is still beneficial even if it comes online after word learning has begun (see 4.4.3).
3. The feedback for a word learning system can be used as a signal to train the joint attention filter to recognise *when* it is appropriate to filter speech (see 5.6.3).

To summarise, a JAF will help accelerate word learning and word learning will train a JAF. Both skills develop over time and experimental evidence reviewed in Chapter 2 suggests that they develop in parallel, primarily during the second year of life. This relationship creates a bootstrapping problem. How can I simulate the two parts of my model developing at the same time when they have previously been used to train each other? The answer is explored in the following experiment.

5.7.2 Methods

The approach of the previous experiment will be repeated, with one major change: the SWLMs are not pre-trained on the vocabulary. I will now describe the algorithm used for training the combined network. The process begins in a manner identical

to the previous experiment: an object-action pair from the training set is fed into the network as input; activation propagates throughout the network until it settles; one node is chosen as the winner in the gating layer; if the SWLMs are activated, a reinforcement signal is generated. Connection weights between the PFC and the gating layer are adjusted as in the previous experiment.

The difference in the algorithm lies in the training of the SWLMs. The information that is let through the filter will vary greatly from epoch to epoch, especially early in training. Because of the inherent random exploration, caused by the initial weights and a lack of prior knowledge, the training examples for which the SWLMs are enabled could theoretically have *no* overlap between the first few epochs. There are difficulties training a backprop network on individual, isolated examples. Many neural network implementations, including supervised-backpropagation, are susceptible to what is known as **catastrophic forgetting**. That is, attempts to learn new patterns lead to an interference with previously-learned patterns. This interference can cause the original patterns to be forgotten unless all of the patterns—new and old—are trained together using numerous examples (see also Frean and Robins, 1999).

The retraining of previously-learned inputs is known as **rehearsal** in the field of neural networks. There is some evidence (e.g., Gershkoff-Stowe and Smith, 1997) that children experiencing a rapid vocabulary growth in the second year of life will demonstrate a transient period of naming errors, possibly caused by an interference between new and previously-learned words; some sort of rehearsal solution is not unwarranted. I will not attempt to implement a biologically-plausible rehearsal algorithm at this time. Instead, the network maintains a sliding window of the recent training examples which have been allowed into the SWLMs by the gating layer. Each object-action pair that the filter lets through is placed in a buffer. When the buffer reaches a certain size, M , the inputs stored up until that point are used as training inputs to the SWLMs. The contents of the buffer, plus R rehearsals of the previously-learned words, are considered one ‘epoch’ for the purpose of batch training an SWLM; weights are updated after one such epoch, consistent with the batch training used in my previous experiments. After one pass through the buffer, the pair that has been in the buffer the longest is discarded. (The buffer is implemented as a first-in-first-out queue, so it is a simple matter to remove the head item.) Thus, the buffer will always store the M most recent conceptual pairs that have been let through the JAF.

The network parameters used for this experiment are shown in Table 5.4. It is important to reiterate that the SWLMs have no prior knowledge when training begins.

Network Parameter	Value
$learn_{default}$	0.03
$learn_{reward}$	0.15
$learn_{punish}$	-0.02
SWLM Learning Constant	0.2
SWLM Hidden Units	2 * vocabulary size
Rehearsals per epoch, R	0.25 * size of training set / size of vocabulary
Size of SWLM training buffer, M	0.15 * size of training set

Table 5.4: Network parameters used in Experiment F.

As such, initially they will not map any words correctly and will be generating punishment signals almost 100% of the time. Because of this fact, the magnitude of the punishment must be set very small. Such an implementation is justifiable, however, on the grounds that infants receive very little explicit negative feedback about their word use (Rohde and Plaut, 1999). If the parameters are set too low, of course, training will take too long.

The network was tested on a training set of 10 nouns and 10 verbs, generated using a CDS of 0.10. Trial runs showed that the network had a tendency to plateau and not converge to an error term less than 0.10. As such, an extra exit condition was added to cease the run if the error had not decreased in the last 100 epochs.

5.7.3 Results

Results, summarised in Figure 5.13, were mixed. The network never learned the entire vocabulary, suggesting my exit terms might have been too aggressive; overall, 80% of the vocabulary was learned. MOTHER-TALK was perfectly learned as a predictor of reward in all runs, after about 100 epochs. The accuracy for turning off the SWLMs in other circumstances seemed to peak around 90%.

5.7.4 Discussion

Although the performance never reached 100% in all measures, the results were very encouraging. The network began to learn with no prior knowledge and—best of all—the filter and the SWLM learned in parallel. That is, the simulation suggests that learning to recognise communicative intentions and word acquisition could develop in

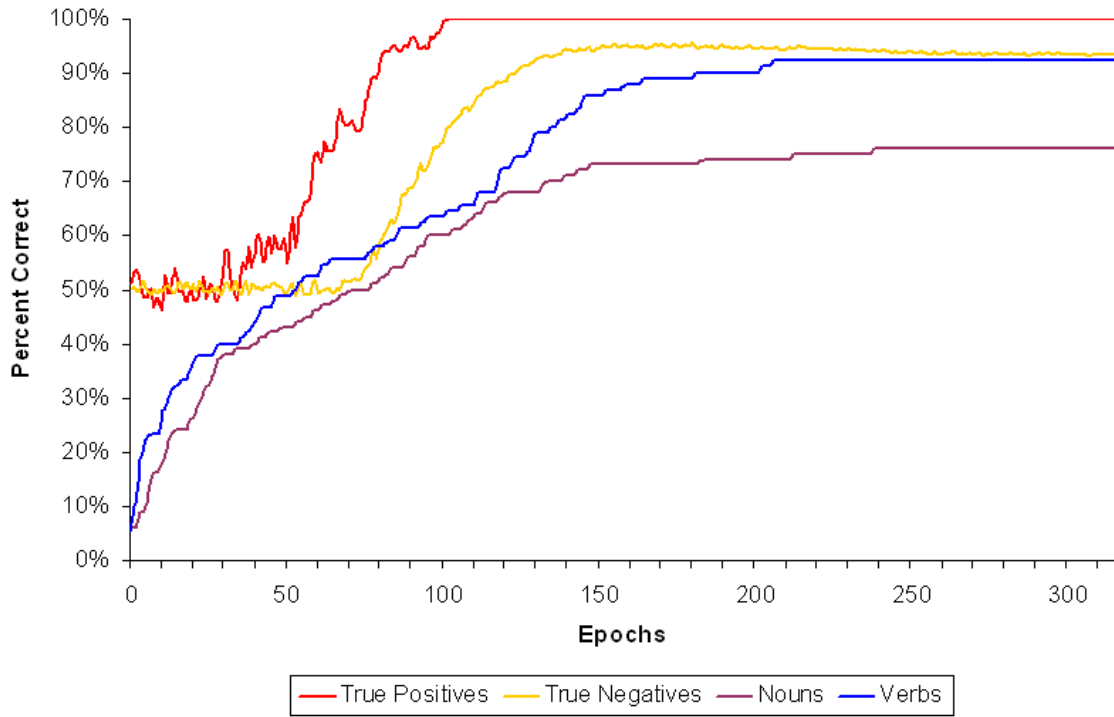


Figure 5.13: Mean results for Experiment F.

parallel, interacting in a manner that is beneficial to each process.

Preliminary testing with other vocabulary sizes (10 and 30 words) produced even more varied results, suggesting that I might not yet have found the network parameters that generalise well to all data sets. However, there is no inherent reason why the solution should not scale. Finding the optimal parameters for the network remains a topic for future work.

As I speculated earlier, it is possible that the aid of joint attention is not essential for the very first words, but merely acts as an accelerator to later word learning. If this is the case, perhaps I should experiment with giving the SWLMs a ‘head start’ and not begin to train the filter until after they have learned a couple of words. Alternatively, the learning constants for the reinforcement algorithm could vary over time, starting small and getting larger as the infant becomes more aware of his environment.

Chapter 6

Conclusion

In this thesis, I have discussed the results from a number of experiments run using my computational models of language learning. In Chapter 4, I explored the idea that infants have a selective ‘word learning mode’. My simulations proved the value of being able to recognise the communicative intentions described by the social-pragmatic theory of language learning (Akhtar and Tomasello, 2000). I showed that, if these intentions can be recognised, word learning mode can be selectively enabled in a way which improves the efficiency of word learning. In Chapter 5, I appealed to evidence from cognitive neuroscience to propose a neural basis for the recognition of communicative intentions. I also proposed a method for the process of *learning* when to enable word learning mode—i.e., to learn the significance of recognising communicative intentions when they take the form of mother-initiated joint attention—and demonstrated how such a process could work.

6.1 Evaluation

One way to evaluate my model is based on the same criteria that I discussed in Chapter 3. To recap, these criteria were: bootstrapping, acceleration, noise, rapid mapping, development, difficulty, and neurological plausibility.

My model clearly satisfies the bootstrapping constraint, as it learns its first words with no prior knowledge; moreover, I also demonstrated how the JAF could learn when to turn on through its own bootstrapping.

My model also easily meets the criterion of acceleration, which specifies that the rate of vocabulary acquisition varies throughout development. In 4.4.3, word learning accelerated when the JAF was activated partway through the learning process. Exper-

iment F (see 5.7.3) also provided evidence that this effect can still be seen if the ability to identify communicative situations is developed in parallel to the word learning.

Noise is handled very well by my model, reflecting the many distractions that an infant will encounter in everyday life. Some of my simulations included a much higher percentage of noise than was reported for any of the models discussed in Chapter 3. Studies like the one by Harris *et al.* (1986) suggest that this parameter should be set very high indeed.

The criterion of development suggests that the structure of a network should evolve throughout training, in order to reflect the rapid changes that occur in the infant brain as it develops. The requirement set out for this constraint was that the network architecture change in some way over the course of training, beyond a simple adjustment of connection strengths between layers. I have satisfied this constraint by adding the gating functionality and having it learn independently of the supervised training of the word subnetwork.

Whether or not my network models rapid mapping is not clear. Children can sometimes learn words in a single exposure, especially after 24 months of age. If an epoch is considered to be an ‘exposure’, a cursory examination of my experimental results does find instances where one or more words were learned in a single epoch. However, equating an entire epoch with an exposure is *not* realistic as each epoch consists of multiple instances of different words. The architecture of a backprop network simply does not lend itself well to single-exposure learning. The learning algorithm is one of gradient descent, guaranteeing monotonically-decreasing error in the system. It is designed to prioritise successful convergence over speed. It is important to note, however, that my conceptual model is very friendly to the notion of rapid mapping. Indeed, one could surmise that focusing on communicative situations would scaffold the process of rapid mapping. In order to realise such an effect in my model, however, I would possibly need to alter the learning regime used by the SWLMs.

Of the criteria, mentioned above, the one I have satisfied the least concerns word difficulty. Because of the manner in which my network inputs are encoded, they are all of theoretically equal difficulty to learn. In practice, of course, this is not exactly true. Random initial weights will inevitably bias some input patterns to be learned more easily. My randomly-generated training sets would also cause some words to be represented with higher frequency than others. However, none of these parameters were *controlled* to approach levels that would accurately reflect the relative difficulties of learning words.

There is one significant item remaining for discussion. I have demonstrated the possible value that recognising communicative intentions can have to facilitate word learning. However, my realisation of this process hinges on the control of a word learning mechanism. That is, my argument proceeds from the assumption that an infant can turn a word learning mode ‘on’ or ‘off’. I have cited no direct evidence of this assumption, because I am not aware of any explicit proof of such a mechanism. This thesis is intended to investigate the plausibility of the mechanism from a computational perspective. There is increasing evidence, however, that the brain might have different modes for different sorts of processing. One argument for this position is my discussion in Chapter 5 about an integrated action-execution/action-recognition system and how it might operate via a mode-switching mechanism.

Experiments like those of Baldwin (1991) definitely demonstrate that there is something *selective* about an infant’s word learning process. They will learn a word under some circumstances and not under others. Those who do not subscribe to social-pragmatic theory would argue any number of other reasons for why the word is not learned in some cases including saliency, attention, memory, etc. I believe that many of the studies I have cited discredit these alternative explanations. Infants do not merely learn words for the most salient objects (Baldwin, 1993; Houston-Price *et al.*, 2006). Suggesting that memory is the only factor is an even weaker argument, because it would require that all the bad examples are conveniently forgotten and only a subset of the good ones are learned. Even those researchers that argue for attention-based learning (e.g., Regier, 2005) do so based on a *selective* process. Therefore, it seems safe to argue that the process of word learning is selective in some manner, whether the selection process is controlled by intention recognition or some other influence. I hope to see experimental research in this area that would begin to clear up some of the mystery. The challenge, then, is to develop an experimental procedure that could look for evidence of a switch in neural activity when infants engage in a word learning situation.

I propose looking for this evidence using a neuroimaging technique. fMRI studies are quite ubiquitous at the moment. However, fMRI has poor temporal resolution; an infant switching in and out of a word learning mode would likely need to do so in less than a second. Enticing a 12- to 24-month-old to remain stationary in an fMRI scanner would also be problematic. However, EEG has better temporal resolution and has already been employed in studies with infants at least as young as 6 months of age (e.g., Benasich, Choudhury, Friedman, Realpe-Bonilla, *et al.*, 2006). Therefore, my

proposal is to record EEG during a study similar to those of Baldwin *et al.* (1996), Woodward and Hoyne (1999), or Akhtar *et al.* (2001). It is difficult to know at the time whether an infant is actually learning a word or not. After the testing phase, however, a comparison of EEG activity could be made between those situations where word learning occurred and when it did not. I can only speculate about what sort of difference would be indicative of the sort of mode switch I have hypothesised. Based on my model, however, it would likely be related to activity in the frontal part of the brain.

6.2 Future Work

There are a few obvious extensions that could be made to improve the model I have presented in this thesis. One future step would be to optimise the PCN (see Appendix A) and integrate it with the intention recognition word learning network. The simulations should also be scaled up to verify that the results hold for the full 300-word vocabulary of an average 24-month-old (Ganger and Brent, 2004). In that case, the recognition of communicative intentions—as well as any vocabulary spurt—should occur when the vocabulary reaches around 50 words, consistent with an 18-month-old.

Speaking of the vocabulary spurt, it would be interesting to apply Ganger and Brent’s full technique to the data from my experiments, permitting a quantitative analysis of what percentage of the results produce a clear point of inflection in the rate of word learning. Tweaking various parameters of the network might control the occurrence of a spurt, suggesting a prediction of individual differences.

I believe it would also be beneficial to consider a network that is more dynamic throughout the training period. As I have stated repeatedly, an infant brain develops dramatically throughout the second year of life. I would like to implement some sort of scheme for having the network structure develop more dramatically as the infant develops (see also Elman, 1990). Perhaps the number of nodes in the network could be changed as the vocabulary increases.

In terms of neurological plausibility, the rehearsal process used in Section 5.7.3 is problematic. It assumes that the network retains prior information by rehearsing it, requiring access to the original training information. By analogy, this would require the infant’s brain to remember its vocabulary by practicing the words it already knows. This solution is rather circular in the sense that he remembers what he knows by remembering what he knows. Robins (1996) argued that ‘pseudorehearsal’ of neural

networks is a more plausible alternative. Pseudorehearsal involves feeding a random set of inputs into the network; the resulting output is later used as a rehearsal target. Thus, pseudorehearsal is like rehearsal but it does not require access to the original inputs on which the network was trained. It only requires a sense of the network's state, a snapshot of synaptic connections in a sense, before the new material was introduced. Robins proposed that pseudorehearsal is analogous to the long-term memory consolidation process that is believed to happen in the mammalian brain during sleep. Therefore, my model could be improved by using pseudorehearsal in place of the current rehearsal solution.

Even once the plausibility of the joint attention filter is optimised, the fact remains that the word learning subnetwork is being trained using supervised learning. As I have already mentioned, such a form of learning is also neurobiologically improbable; infants do not learn language solely in a highly-structured environment with constant and accurate feedback. Such learning also has difficulties simulating rapid mapping. Other improvements to my model could be realised by investigating alternatives to the supervised backprop network. Interesting work using other architectures, like **self-organising maps** (Li *et al.*, 2007), suggests some alternate possibilities.

References

- Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science*, 8(2), 199–209.
- Akhtar, N., Jipson, J., and Callanan, M. A. (2001). Learning Words through Overhearing. *Child Development*, 72(2), 416–430.
- Akhtar, N. and Tomasello, M. (2000). The Social Nature of Words and Word Learning. In *Becoming a Word Learner: A Debate on Lexical Acquisition*, 115–135. Oxford, England: Oxford University Press.
- Altmann, G. T. M. (1997). *The Ascent of Babel: an exploration of language, mind, and understanding*. New York, NY, USA: Oxford University Press.
- Averbeck, B. B., Chafee, M. V., Crowe, D. A., and Georgopoulos, A. P. (2002). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20), 13172–13177.
- Averbeck, B. B. and Lee, D. (2007). Prefrontal Neural Correlates of Memory for Sequences. *Journal of Neuroscience*, 27(9), 2204–2211.
- Averbeck, B. B., Sohn, J.-W., and Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature Neuroscience*, 9(2), 276–281.
- Baddeley, A., Gathercole, S., and Papagno, C. (1998). The Phonological Loop as a Language Learning Device. *Psychological Review*, 105(1), 158–173.
- Baddeley, A. D. and Hitch, G. J. (1994). Developments in the Concept of Working Memory. *Neuropsychology*, 8(4), 485–493.
- Baldwin, D. A. (1991). Infants’ Contribution to the Achievement of Joint Reference. *Child Development*, 62, 875–890.

- Baldwin, D. A. (1993). Early Referential Understanding: Infants' Ability to Recognize Referential Acts for What They Are. *Developmental Psychology*, 29(5), 832–843.
- Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., *et al.* (1996). Infants' Reliance on a Social Criterion for Establishing Word-Object Relations. *Child Development*, 67, 3135–3153.
- Baldwin, D. A. and Moses, L. J. (1996). The Ontogeny of Social Information Gathering. *Child Development*, 67, 1915–1939.
- Baron-Cohen, S. (1989). Perceptual role taking and protodeclarative pointing in autism. *British Journal of Developmental Psychology*, 7, 113–127.
- Baron-Cohen, S. (1995). *Mindblindness: an essay on autism and theory of mind*. Cambridge, MA, USA: MIT Press.
- Bates, E., Thal, D., Whitesell, K., *et al.* (1989). Integrating Language and Gesture in Infancy. *Developmental Psychology*, 25(6), 1004–1019.
- Benasich, A. A., Choudhury, N., Friedman, J. T., Realpe-Bonilla, T., *et al.* (2006). The infant as a prelinguistic model for language learning impairments: Predicting from event-related potentials to behavior. *Neuropsychologia*, 44, 396–411.
- Binkofski, F., Buccino, G., Posse, S., Seitz, R. J., *et al.* (2000). A fronto-parietal circuit for object manipulation in man: Evidence from an fmri-study. *European Journal of Neuroscience*, 11(9), 3276–3286.
- Bloom, L. (2000). Pushing the Limits on Theories of Word Learning. *Monographs of the Society for Research in Child Development*, 65(3), 124–135.
- Braver, T. S. and Cohen, J. D. (2000). On the Control of Control: The Role of Dopamine in Regulating Prefrontal Function and Working Memory. In *Control of Cognitive Processes: Attention and Performance*, Volume XVIII, Cambridge, MA, USA, 713–737. MIT Press.
- Burnham, D. and Dodd, B. (2004). Auditory-Visual Speech Integration by Prelinguistic Infants: Perception of an Emergent Consonant in the McGurk Effect. *Developmental Psychobiology*, 45(4), 204–220.
- Butterworth, G. and Harris, M. (1994). *Principles of developmental psychology*. Hove, UK: Lawrence Erlbaum.

- Butterworth, G. and Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9, 55–72.
- Carpenter, M., Nagell, K., and Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, 63(4), 1–174.
- Caza, G. A. (2007). Learning to recognise relevant ambient speech: a computational model. In P. Howie and D. Livesey (Eds.), *Proceedings of 15th Biennial Conference of the Australasian Human Development Association*, Volume 59 of *Australian Journal of Psychology*, 156. Australian Psychological Society.
- Cooper, R. and Shallice, T. (2000). Contention Scheduling and the Control of Routine Activities. *Cognitive Neuropsychology*, 17(4), 297–338.
- Corina, D. P., Gibson, E. K., Martin, R., Poliakov, A., *et al.* (2004). Dissociation of Action and Object Naming: Evidence From Cortical Stimulation Mapping. *Human Brain Mapping*, 24, 1–10.
- Damasio, A. R. (1995). On Some Functions of the Human Prefrontal Cortex. *Annals of the New York Academy of Sciences*, 769, 241–251.
- Damasio, A. R. and Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences of the United States of America*, 90, 4957–4960.
- Damasio, H., Grabowski, T. J., Tranel, D., *et al.* (1996). A neural basis for lexical retrieval. *Nature*, 380, 499–505.
- Daniele, A., Giustolisi, L., Silveri, M. C., *et al.* (1994). Evidence for a Possible Neuroanatomical Basis for Lexical Processing of Nouns and Verbs. *Neuropsychologia*, 32(11), 1325–1341.
- di Pellegrino, G., Fadiga, L., Fogassi, L., *et al.* (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research*, 91(1), 176–180.
- Diana, R. A., Yonelinas, A. P., and Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences*, 11(9), 379–386.

- Dorris, M. C., Klein, R. M., Everling, S., and Munoz, D. P. (2002). Contribution of the Primate Superior Colliculus to Inhibition of Return. *Journal of Cognitive Neuroscience*, 14(8), 1256–1263.
- Dreher, J.-C. and Berman, K. F. (2002). Fractionating the neural substrate of cognitive control processes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22), 14595–14600.
- Dubois, B., Levy, R., Verin, M., Teixeira, C., *et al.* (1995). Experimental Approach to Prefrontal Functions in Humans. *Annals of the New York Academy of Sciences*, 769, 41–60.
- Dunbar, K. and Sussman, D. (1995). Toward a Cognitive Account of Frontal Lobe Function: Simulation Frontal Lobe Deficits in Normal Subjects. *Annals of the New York Academy of Sciences*, 769, 289–304.
- Eichenbaum, H., Yonelinas, A. P., and Ranganath, C. (2007). The Medial Temporal Lobe and Recognition Memory. *Annual Review of Neuroscience*, 30, 123–152.
- Eilan, N. (2005). Joint Attention, Communication, and Mind. In N. Eilan, C. Hoerl, T. McCormack, and J. Roessler (Eds.), *Joint Attention: Communication and Other Minds*, 1–33. Oxford, England: Oxford University Press.
- 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.
- Fadiga, L., Fogassi, L., Pavesi, G., and Rizzolatti, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology*, 73(6), 2608–2611.
- Falck-Ytter, T., Gredebäck, G., and von Hofsten, C. (2006). Infants predict other people’s action goals. *Nature Neuroscience*, 9(7), 878–879.
- Fernald, A. and Morikawa, H. (1993). Common Themes and Cultural Variations in Japanese and American Mothers’ Speech to Infants. *Child Development*, 64, 637–656.
- Flanagan, J. R. and Johansson, R. S. (2003). Action plans used in action observation. *Nature*, 424, 769–771.

- Floor, P. and Akhtar, N. (2006). Can 18-Month-Old Infants Learn Words by Listening In on Conversations? *Infancy*, 9(3), 327–339.
- Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., *et al.* (2005). Parietal Lobe: From Action Organization to Intention Understanding. *Science*, 308, 662–667.
- Frean, M. and Robins, A. (1999). Catastrophic forgetting in simple networks: an analysis of the pseudorehearsal solution. *Computational Neural Systems*, 10, 2276–236.
- Fuster, J. M. and Alexander, G. E. (1971). Neuron Activity Related to Short-Term Memory. *Science*, 173(3997), 652–654.
- Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2), 593–609.
- Gallese, V. and Lakoff, G. (2005). The Brain’s Concepts: The Role of the Sensory-Motor System in Conceptual Knowledge. *Cognitive Neuropsychology*, 22(3/4), 455–479.
- Ganger, J. and Brent, M. R. (2004). Reexamining the Vocabulary Spurt. *Developmental Psychology*, 40(4), 621–632.
- Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences*, 3(11), 410–419.
- Gershkoff-Stowe, L. and Smith, L. B. (1997). A Curvilinear Trend in Naming Errors as a Function of Early Vocabulary Growth. *Cognitive Psychology*, 34, 37–71.
- Ghazanfar, A. A. (2008). Language evolution: neural differences that make a difference. *Nature Neuroscience*, 11(4), 382–384.
- Hare, B., Brown, M., Williamson, C., and Tomasello, M. (2002). The Domestication of Social Cognition in Dogs. *Science*, 298, 1634–1636.
- Hari, R., Forss, N., Avikainen, S., Kirveskari, E., *et al.* (1998). Activation of human primary motor cortex during action observation: A neuromagnetic study. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 15061–15065.
- Harris, M., Jones, D., Brookes, S., and Grant, J. (1986). Relations between the non-verbal context of maternal speech and rate of language development. *British Journal of Developmental Psychology*, 34, 261–268.

- Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. New York, NY, USA: Macmillan.
- Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Redwood City, CA, USA: Addison-Wesley.
- Hoehl, S., Reid, V., Mooney, J., and Striano, T. (2007). What are you looking at? Infants’ neural processing of an adult’s object-directed gaze. *Developmental Science*, 11, 10–16.
- Hood, B. M., Willen, D. J., and Driver, J. (1998). Adult’s Eyes Trigger Shifts of Visual Attention in Human Infants. *Psychological Science*, 9(2), 131–134.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79, 2554–2558.
- Houghton, G. and Hartley, T. (1995). Parallel Models of Serial Behaviour: Lashley Revisited. *Psyche*, 2(25), 1–25.
- Houston-Price, C., Plunkett, K., and Duffy, H. (2006). The use of social and salience cues in early word learning. *Journal of Experimental Psychology*, 95, 27–55.
- Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., *et al.* (1999). Cortical Mechanisms of Human Imitation. *Science*, 286(5449), 2526–2528.
- Kaminski, J., Riedel, J., Call, J., and Tomasello, M. (2005). Domestic goats, *Capra hircus*, follow gaze direction and use social cues in an object choice task. *Animal Behaviour*, 69, 11–18.
- Kerns, J. G., Cohen, J. D., Stenger, V. A., and Carter, C. S. (2004). Prefrontal Cortex Guides Context-Appropriate Responding during Language Production. *Neuron*, 43, 283–291.
- Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, 4(4), 138–147.
- Knott, A. (2008). Sensorimotor cognition and natural language syntax. In preparation.
- Langacker, R. W. (1987). Nouns and Verbs. *Language*, 63(1), 53–94.
- Li, P., Zhao, X., and MacWhinney, B. (2007). Dynamic Self-Organization and Early Lexical Development in Children. *Cognitive Science*, 31, 581–612.

- Liberman, A. M. and Mattlingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1–36.
- Marino, L., Connor, R. C., Fordyce, R. E., Herman, L. M., *et al.* (2007). Cetaceans Have Complex Brains for Social Cognition. *PLoS Biology*, 5(5), 966–972.
- Mayr, U. and Keele, S. W. (2000). Changing Internal Constraints on Action: The Role of Backward Inhibition. *Journal of Experimental Psychology*, 129(1), 4–26.
- McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
- McMurray, B. (2007). Defusing the Childhood Vocabulary Explosion. *Science*, 317, 631.
- Meltzoff, A. N. (1995). Understanding the Intentions of Others: Re-Enactment of Intended Acts by 18-Month-Old Children. *Developmental Psychology*, 31(5), 838–850.
- Miller, E. K. (1999). The Prefrontal Cortex: Complex Neural Properties for Complex Behavior. *Neuron*, 22(1), 15–17.
- Miller, E. K. and Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, 24, 167–202.
- Miller, E. K., Erickson, C. A., and Desimone, R. (1996). Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque. *Journal of Neuroscience*, 16(16), 5154–5167.
- Moore, C., Angelopoulos, M., and Bennett, P. (1999). Word Learning in the Context of Referential and Salience Cues. *Developmental Psychology*, 35(1), 60–68.
- Mundy, P., Block, J., Delgado, C., Pomares, Y., *et al.* (2007). Individual Differences and the Development of Joint Attention in Infancy. *Child Development*, 78(3), 938–954.
- Mundy, P. and Newell, L. (2007). Attention, Joint Attention, and Social Cognition. *Current Directions in Psychological Science*, 16(5), 269–274.
- Muthukumaraswamy, S. D., Johnson, B. W., and McNair, N. A. (2004). Mu rhythm modulation during observation of an object-directed grasp. *Cognitive Brain Research*, 19, 195–201.

- Newport, E. L. (1990). Maturational Constraints on Language Learning. *Cognitive Science*, 14, 11–28.
- O’Grady, W. (2005). *How Children Learn Language*. Cambridge, UK: Cambridge University Press.
- Pascual-Leone, A., Grafman, J., and Hallett, M. (1995). Procedural Learning and Prefrontal Cortex. *Annals of the New York Academy of Sciences*, 769, 61–70.
- Pfurtscheller, G., Neuper, C., and Krausz, G. (2000). Functional dissociation of lower and upper frequency mu rhythms in relation to voluntary limb movement. *Clinical Neurophysiology*, 111(10), 1873–1879.
- Pulvermüller, F., Shtyrov, Y., and Ilmoniemi, R. (2005). Brain Signatures of Meaning Acces sin Action Word Recognition. *Journal of Cognitive Neuroscience*, 17(6), 884–892.
- Regier, T. (2005). The Emergence of Words: Attentional Learning in Form and Meaning. *Cognitive Science*, 29, 819–865.
- Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research*, 3(2), 131–141.
- Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., *et al.* (1996). Localization of grasp representations in humans by PET .1. Observation versus execution. *Experimental Brain Research*, 111(2), 246–252.
- Robins, A. (1996). Consolidation in neural networks and in the sleeping brain. *Connection Science*, 8, 259–275.
- Rohde, D. L. T. and Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, 72, 67–109.
- Sabbagh, M. A. and Baldwin, D. A. (2005). Understanding the Role of Communicative Intentions in Word Learning. In N. Eilan, C. Hoerl, T. McCormack, and J. Roessler (Eds.), *Joint Attention: Communication and Other Minds*, 165–184. Oxford, England: Oxford University Press.
- Saffran, J. R. (2003). Statistical Language Learning: Mechanisms and Constraints. *Current Directions in Psychological Science*, 12(4), 110–113.

- Saffran, J. R., Aslin, R. N., and Newport, E. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274, 1926–1928.
- Samuelson, L. K. and Smith, L. B. (1998). Memory and Attention Make Smart Word Learning: An Alternative Account of Akhtar, Carpenter, and Tomasello. *Child Development*, 69(1), 94–104.
- Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, 80(6), 1–27.
- Semendeferi, K., Lu, A., Schenker, N., and Damasio, H. (2002). Humans and great apes share a large frontal cortex. *Nature Neuroscience*, 5(3), 272–276.
- Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge, MA, USA: Cambridge University Press.
- Sirigu, A., Zalla, T., Pillon, B., Grafman, J., *et al.* (1995). Planning and Script Analysis following Prefrontal Lobe Lesions. *Annals of the New York Academy of Sciences*, 769, 277–288.
- Siskind, J. M. (1994). Lexical Acquisition in the Presence of Noise and Homonymy. *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI)*, 760–766.
- Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61, 39–91.
- Srimal, R. and Curtis, C. E. (2008). Persistent neural activity during the maintenance of spatial position in working memory. *NeuroImage*, 39, 455–468.
- Striedter, G. F. (2005). *Principles of Brain Evolution*. Sunderland, MA, USA: Sinauer Associates.
- Swingler, K. (1996). *Applying Neural Networks: A Practical Guide*. London, UK: Academic Press.
- Thoermer, C. and Sodian, B. (2001). Preverbal infants’ understanding of referential gestures. *First Language*, 21, 245–264.
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA, USA: Harvard University Press.

- Tomasello, M. and Akhtar, N. (2003). What paradox? A response to Naigles (2002). *Cognition*, 88, 317–323.
- Tomasello, M., Call, J., and Hare, B. (2003). Chimpanzees understand psychological states - the question is which ones and to what extent. *Trends in Cognitive Science*, 7(4), 153–156.
- Triesch, J., Teuscher, C., Deák, G. O., and Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, 9(2), 125–157.
- Tyler, L. K. and Marslen-Wilson, W. (2008). Fronto-temporal brain systems supporting spoken language comprehension. *Philosophical Transactions of the Royal Society B*, 363, 1037–1054.
- van Oijen, J. (2006). A model of the relationship between language and sensorimotor cognition. <http://www.cs.otago.ac.nz/research/techreports.html>.
- Vigneau, M., Beaucousin, V., Hervé, P., Duffau, H., *et al.* (2006). Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *NeuroImage*, 30, 1414–1432.
- Webb, A., Knott, A., and MacAskill, M. (2008). Eye movements during transitive action observation have sequential structure. In press.
- Woodward, A. L. (1998). Infants selectively encode the goal object of an actor’s reach. *Cognition*, 69, 1–34.
- Woodward, A. L. (2000). Constraining the Problem Space in Early Word Learning. In *Becoming a Word Learner: A Debate on Lexical Acquisition*, 81–113. Oxford, England: Oxford University Press.
- Woodward, A. L. (2003). Infants’ developing understanding of the link between looker and object. *Developmental Science*, 6, 297–311.
- Woodward, A. L. and Hoyne, K. L. (1999). Infants’ Learning about Words and Sounds in Relation to Objects. *Child Development*, 70(1), 65–77.
- Yu, C. and Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70, 2149–2165.
- Zanini, S. (2008). Generalised script sequencing deficits following frontal lobe lesions. *Cortex*, 44, 140–149.

Zanini, S., Rumiati, R. I., and Shallice, T. (2002). Action Sequencing Deficit Following Frontal Lobe Lesion. *Neurocase*, 8, 88–99.

Appendix A

A Model of Plan Competition

As mentioned in Chapter 5, the human PFC is believed to be important to intention-guided behaviour, as well as to working memory. I argued that the PFC maintains a representation of an entire proposition, describing an intention. This representation could be of one's own intention, recognition of another's intention, or even a neural replay of a past intention. My proposal in Chapter 4 was that children are able to switch in and out of a word learning *mode*, when it is appropriate. Recognition of a communicative intention is one of the signals that an appropriate word learning situation is about to be presented. I modelled this by having the infant recognise the event MOTHER-TALK and use it as a trigger to engage with the mother in joint attention. The word learning situation is completed by following her gaze and associating what she is looking at with the word she utters. However, the process essentially involves the representation of *two* propositions. For example, if the mother is talking about a dog jumping, there is MOTHER-TALK *and* DOG-JUMP. Given that working memory is limited (Gathercole, 1999), this leads to a real-estate problem.

I will now discuss a theoretical solution to this problem. When the infant recognises a communicative intention, the contents of his working memory (i.e., the proposition in the PFC) must switch from an action-observation representation (MOTHER-TALK) to a representation of the sensorimotor concepts of the proposition described by the word he is trying to learn. Previously, I have been ambiguous about the process for such a switch. It was implicitly assumed to occur at the appropriate time. Now, I shall propose a mechanism for plan switching, implemented using a computational model. I will refer to it as the **Plan Competition Network (PCN)**. It is my argument that my PCN—in many ways an extension of Guided Activation Theory (Miller and Cohen, 2001)—is generalisable to switching between plans in many contexts. However, I am

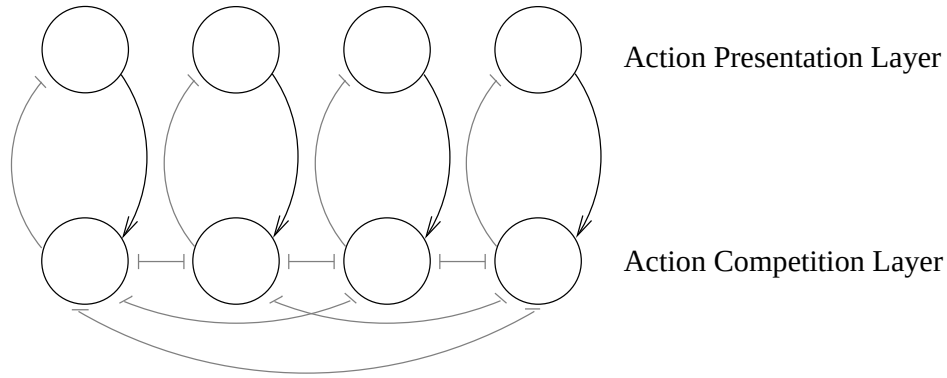


Figure A.1: A competitive queuing model that demonstrates contention scheduling (based on Houghton and Hartley, 1995). Arrow-headed connections are excitatory; other connections are inhibitory.

primarily interested in its applicability to the recognition of communicative intentions and, where possible, I will relate it to my word learning model of intention recognition.

A.1 Background

A.1.1 Serial Behaviour

As mentioned in Chapter 5, stimulus-response mappings are not completely deterministic for healthy adults. There are multiple possible responses to the same stimulus, and multiple stimuli may be presented concurrently. However, it is often difficult to perform more than one action simultaneously. They often must be executed serially. How is a choice made between competing responses, and how are they ordered? **Contention scheduling** networks have been used to model serially-ordered behaviour. The term is used to refer to “routine selection between routine actions” (Shallice, 1988, p. 334). Different actions are assumed to present themselves in parallel and be in conflict to be executed, hence the use of the term ‘contention’. Selection of the appropriate action is considered to be the task performance bottleneck (Cooper and Shallice, 2000). One model of contention scheduling, which Houghton and Hartley (1995) dubbed **competitive queuing (CQ)**, is illustrated in Figure A.1.

The *presentation layer* represents the action schemata, consisting of tasks to be completed serially. The *competition layer* acts as a ‘competitive filter’. The lateral inhibition in the latter enforces a ‘winner-take-all’ decision, with one neuron becoming activated while simultaneously inhibiting its competitors. The implementation of this

inhibition through the lateral weights, rather than an external source of inhibition, is elegant because it allows the competing schemata to have varying degrees of overlap and influence on each other (Cooper and Shallice, 2000). Once a winner is declared, the selected response inhibits itself through its connections back to the presentation layer. Thus, if a gradient of activation is provided as input, the strongest response will be chosen. But, after that response inhibits itself, the next strongest will rise to the top, and so on. It is labelled a ‘queuing’ model because the actions are performed one after the other without an ordered structure, analogous to “a crowded bar with only one bartender” (Houghton and Hartley, 1995).

CQ is useful for simulating the execution order of short, ballistic actions. One such example is typing (as reviewed in Houghton and Hartley, 1995). It has been shown to successfully model not only correct spelling behaviour, but also common errors such as reversals (e.g., “teh” rather than “the”) and distorted repetitions (e.g., “aatend” rather than “attend”). Moreover, CQ models have been implemented with unsupervised, Hebbian learning algorithms (Houghton and Hartley, 1995), which are more neurologically plausible than a supervised learning approach.

Neurological evidence that CQ is indeed a realistic model of what occurs in the monkey PFC comes from a study discussed in Chapter 5 (Averbeck *et al.*, 2002). Rhesus macaques were trained to draw sequential line segments to produce geometric shapes and single-cell recording revealed that the different movements in the sequence were expressed at different strengths by neuron activity in the PFC. At the beginning of the sequence, all the movements were represented, at strength relative to the expected order of execution; the first movement was represented with strongest activation, followed by the second, etc. Over time, as individual movements were completed, the gradient of activation would continue to represent the serial order. The representation of each segment would reach its maximum strength near the middle of its execution, at which time the representation of the next segment would begin to strengthen.

CQ has a lot of appeal, based on the behavioural and neurophysiological evidence. Such models are too simple, however, to represent action sequences that occur over long or unpredictable courses of time. Also, the discrete units in the competition layer can only represent isolated actions and not multi-step, alternative sequences or plans. Next, I will address the latter problem by discussing a model of associative memory.

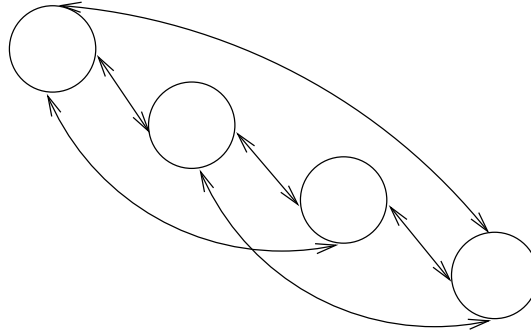


Figure A.2: A Hopfield Net (Hopfield, 1982). All connections are shown as excitatory for the purposes of illustration, but may become inhibitory during training.

A.1.2 Associative Memory

The **Hopfield Net** (Hopfield, 1982) is a well-established neural network model of associative memory. The basic Hopfield Net consists of fully-connected neurons, with no neuron connected back to itself. All connections are bidirectional and the weight for each connection is symmetric. Connection weights begin at zero, and will fluctuate in strength as the network learns. Influence will be either excitatory or inhibitory, depending on whether the weight is currently positive or negative, and it is possible for a weight to change sign multiple times during training. Figure A.2 shows an example of a small Hopfield Net.

Each node has an external input (not shown in Figure A.2.) During training, this input value is fed into an activation function which is used to set the output value of the node; connection weights are adjusted accordingly. During testing, each output is ‘soft-clamped’, or initialized, to the value of its external input. All nodes are then settled, asynchronously and in random order, until the output of each node reflects the activated value of the sum of its inputs.

The network becomes, essentially, a pattern completer; it is attracted to, and tends to converge on, a stable state that consists of a group of associated neurons from the training data. An example is shown in Figure A.3, with a small network trained to recognise two patterns. After training, if the external input for one node was activated in isolation, then the network would settle into a completed version of the learned pattern.

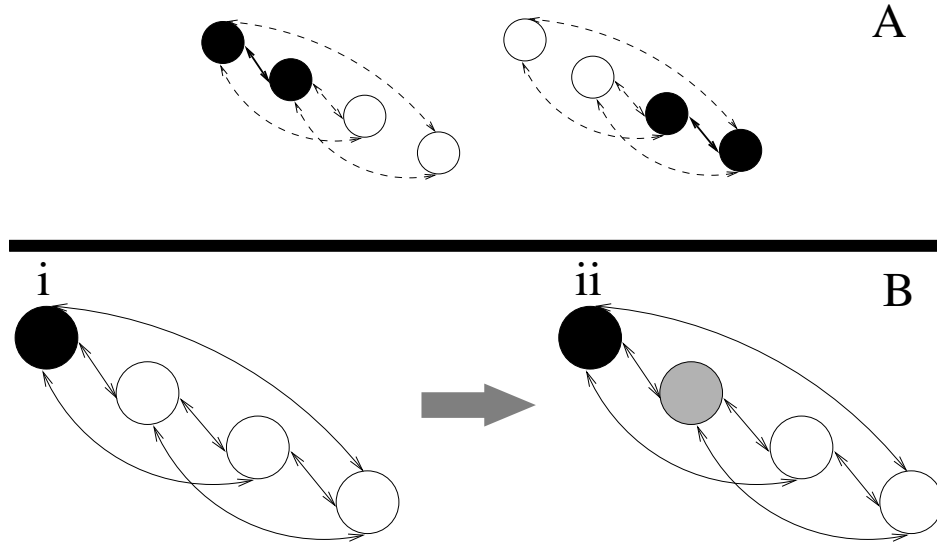


Figure A.3: The expected behaviour of a Hopfield Net. **A:** Two patterns are presented during training; dashed connections are weakened, thick connections are strengthened, and other connections are left unmodified. **B:** After training, (i) a subset of one of the learned patterns is presented and (ii) the network settles into an activation that completes it.

A.1.3 Backward Inhibition

One valuable concept to borrow from competitive queuing is the idea of inhibition back to the presentation layer. There must be a mechanism for clearing the last plan from memory. The phenomenon of self-inhibition in visual attention, referred to as **inhibition of return (IOR)**, is well-established (e.g., Klein, 2000). In visual search experiments, attending to a previously-cued location demonstrates increased reaction times, suggesting a lingering inhibition of attentional resources. Such a restriction makes logical sense from the perspective of search efficiency. If you did not find what you were looking for one location, you should continue your search by looking in new locations rather than reexamining the ones that produced no results. Dorris, Klein, Everling, and Munoz (2002) studied the phenomenon in rhesus monkeys and found evidence that the neural substrate of IOR in primates is the superior colliculus (SC).

Such inhibitory effects do not appear to be limited to the domain of visual search, but are also seen during task switching. Mayr and Keele (2000) performed a behavioural study of **backward inhibition (BI)**, which the authors propose as neces-

sary component of cognitive control. Subjects were instructed to find the object that differed from the others in a display, across one of three different value dimensions. The order in which the dimensions were presented was controlled to create lag-2 repetition and non-repetition cases for analysis; immediate repetitions were not used. Reaction times were significantly longer in repetition conditions than non-repetition conditions. The authors concluded that the cause of the increase was a lingering inhibition of the previously-used approach. Once a task set, or plan, is abandoned, there is an uncontrollable delay before it can be reestablished.

Dreher and Berman (2002) looked for neurological evidence of BI. fMRI was used to scan subjects while they characterised individual letters along one of three different dimensions. BI effects were noted and the researchers found a double dissociation between two different brain regions. The anterior cingulate (ACC) was active when initiating a sequence. The right lateral PFC was active during the phase of residual inhibition before a recently-performed task could be repeated. Thus, the frontal lobe—especially the PFC—is clearly involved in these inhibitory processes.

A.2 Experiment G: The Plan Competition Network

The PCN is a multi-layer neural network to model plan competition in the PFC. The model consists of a *Plan Presentation Layer*, a *Plan Competition Layer*, and a *Plan Termination Gate*. Each of these elements is briefly described below. Figure A.4 shows a simplified version of the network, using a reduced number of nodes.

A.2.1 Design

The Plan Competition Layer is a Hopfield Net, trained to attract to a stable state of associated neurons that represent the winning plan. It is connected to another layer that simulates possible plans, based on current stimuli. Each neuron in this Plan Presentation Layer has a strong forward connection to a neuron in the competition layer, acting as the external input would in a standard Hopfield Net.

This model evolves the backward inhibition property of contention scheduling. Each competition neuron also has an inhibitory connection back to the presentation neuron that is providing it with input. These backward connections are gated by the Plan Termination Gate, meaning it controls the inhibition and only ‘turns it on’ under certain circumstances. (There is neurobiological evidence for gating systems in the PFC (Braver and Cohen, 2000).) It is hypothesized that this gating is used to terminate the

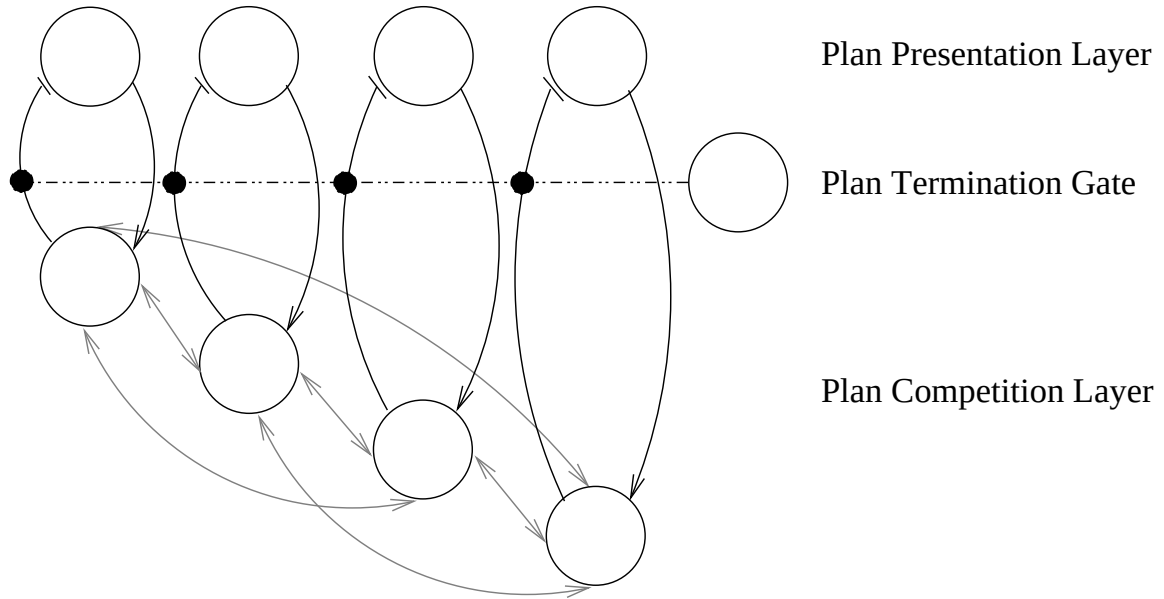


Figure A.4: A simplified representation of the Plan Competition Network. Black connections are of fixed weight, while grey connections vary in strength. The dotted line represents a gate on the inhibitory connections.

current plan and make way for subsequent plans. The simplest reason the current plan could be terminated is because it has been successfully completed. However, there also needs to be cognitive ‘escape plan’ in case the current goal is unachievable or abandoned for other reasons.

The PCN is very similar to the competitive queuing models presented above. One main difference is that the competition layer, represented here by the Hopfield Net, is not one of simple lateral inhibition with an individual neuron ‘winning’ over all competitors. Rather, it is a ‘winning-assembly-takes-all’ system, in which a *group* of associated neurons will become active and inhibit their competitors. In that sense, this network has more in common with later models of contention scheduling that predict the selection of competing action schemata (Cooper and Shallice, 2000).

The other main difference is that the return inhibition to the presentation layer is gated, allowing the plan (i.e., the input) to remain in active memory until the sequence has been completed. A simple, time-delayed backward inhibition is insufficient because of the variability in the amount of time it takes to complete a plan.

A.2.2 Training Parameters

During training, a pattern is presented at the external inputs and propagated forward from the presentation layer to the competition layer. Because all of the flexible connections are symmetric, learning is performed on a by-connection basis, according to the following formula:

$$\Delta w_{ij} = \epsilon y'_i y'_j \lambda . \quad (\text{A.1})$$

The change in weight, w , for the connection between neuron i and neuron j is proportional to the product of the *training* output (see below), y'_n , of each neuron. Here, ϵ represents the learning constant which Hopfield (Hopfield, 1982) set to $\frac{1}{N}$, where N is the total number of neurons. And, λ is a modification of **Oja's Rule** (Hertz, Krogh, and Palmer, 1991), which keeps the weights from growing to infinite magnitude. In this case, λ is the difference between 1 and the absolute value of the weight's current strength. The net effect is to make the values asymptotic as they approach unit magnitude.

Substituting the expanded terms into A.1 yields the following full equation:

$$\Delta w_{ij} = \frac{x_i x_j (1 - |w|)}{N} . \quad (\text{A.2})$$

A scaled training output is used so that neurons that are off at the same time will be associated by having their mutual connections strengthened. The training output is related to the neuron output, y_n , by the following formula:

$$y'_n = 2y_n - 1 . \quad (\text{A.3})$$

Thus, for neurons with discrete activation, the training output A.3 is simply:

$$y'_n = \begin{cases} 1 & \text{if } y_n == 1 \\ -1 & \text{otherwise} \end{cases} . \quad (\text{A.4})$$

Experiments to date have been performed using various sizes of training data, as well as smaller values of the learning parameter, ϵ . It is not yet known if these equations will require modification under a network where the neurons have a continuous activation function.

A.2.3 Preliminary Results

The ideal network would have at least the following characteristics, after training:

1. If there is no input, no neurons will be activated. This case represents a ‘no current plan’ situation.
2. If the current input matches a plan exactly, that plan will win the competition. In the competition layer, the associated neurons will be activated and all others will be off.
3. If the current input is the subset of a plan, the full plan will be activated.
4. If the current input presents multiple plans at equal strength, **one** of the plans will win.
5. If multiple plans are presented as an input gradient, the strongest plan will win.
6. After gating, the last winning plan will be deactivated.
7. If multiple plans are presented and the inhibition is gated, the next strongest plan will take over.

The first three goals have been achieved, as have the last two. For multiple plan situations, the network is currently unpredictable. Most of the time, one plan will win. However, because settling occurs asynchronously in a random order, occasionally a mixture of the two plans will result. The greater the difference in the relative strengths of the two plans, the more likely a clear winner will be chosen. Further research could serve to optimize the behaviour of the Plan Competition network in multiple plan cases.

Initial research consisted of neurons with a discrete, step-based activation function. All neurons in the network are currently moving to a continuous activation function. Further experimentation is planned to determine the most appropriate function.

A.2.4 Discussion

Ultimately, the PCN could be integrated with the word learning network described in Chapter 5, replacing the PFC layer. Just as the word learning network learned when to enable word learning, the PCN should ultimately learn to gate its inhibition at appropriate times.