COGNITIVE MODELING OF SENTENCE MEANING ACQUISITION USING A HYBRID CONNECTIONIST COMPUTATIONAL MODEL INSPIRED BY COGNITIVE GRAMMAR

by

Carson Ka Shing Cheng

B.Sc. (Hons.), University of Alberta, 2005

B.Ed., University of Calgary, 2007

a Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
in the
School of Computing Science
Faculty of Applied Sciences

APPROVAL

Name: Carson Ka Shing Cheng

Degree: Master of Science

Title of Thesis: Cognitive Modeling of Sentence Meaning Acquisition using a Hybrid Connectionist Computational Model inspired by Cognitive Grammar

Examining Committee:

Dr. Arthur Kirkpatrick,
Associate Professor, Computing Science,
Simon Fraser University
Chair

Dr. Robert F. Hadley,
Professor, Computing Science,
Simon Fraser University
Senior Supervisor

Dr. Anoop Sarkar,
Associate Professor, Computing Science,
Simon Fraser University
Supervisor

Dr. Fred Popowich,
Professor, Computing Science,
Simon Fraser University
Examiner

Date Approved: August 3, 2011

Partial Copyright Licence

Abstract

A novel connectionist architecture of artificial neural networks is presented to model the assignment of meaning to test sentences on the basis of learning from *relevantly* sparse input. Training and testing sentences are generated from simple recursive grammars, and once trained, the architecture successfully processes thousands of sentences containing deeply embedded clauses, therefore experimentally showing the architecture exhibits partial semantic and strong systematicities — two properties that humans also satisfy.

The architecture's novelty derives, in part, from analyzing language meaning on the basis of Cognitive semantics (Langacker, 2008), and the concept of affirmative stimulus meaning (Quine, 1960). The architecture demonstrates one possible way of providing a connectionist *processing model* of Cognitive semantics. The architecture is argued to be oriented towards increasing neurobiological and psychological plausibility as well, and will also be argued as being capable of providing an explanation of the aforementioned systematicity properties in humans.

Keywords: Neural Networks; Sentence Meaning Acquisition; Cognitive Semantics; Systematicity; Connectionism; Cognition

To my teachers.

"In the days when Sussman was a novice,
Minsky once came to him as he sat hacking at the PDP-6.
'What are you doing?', asked Minsky.
'I am training a randomly wired neural net to play Tic-Tac-Toe' Sussman replied.
'Why is the net wired randomly?', asked Minsky.
'I do not want it to have any preconceptions of how to play', Sussman said.
Minsky then shut his eyes.
'Why do you close your eyes?', Sussman asked his teacher.
'So that the room will be empty.'
At that moment, Sussman was enlightened."

— *AI Koan*, The Jargon File *(Raymond, 2003)*

Acknowledgments

My parents' unwavering support and encouragement have been key to my success, and I thank them for that and more.

This thesis could never have been completed without the generous support, guidance, and mentorship from my senior supervisor, Dr. Robert F. Hadley; his prompt feedback on my ideas, research, and writing is reflected throughout this thesis as immeasurable improvements. I am grateful to my supervisor, Dr. Anoop Sarkar, and my examiner, Dr. Fred Popowich, for taking the time to provide me with insightful comments that have improved my thesis substantially.

Thanks, also, to Chun-Hing Cheng for his continuing support, encouragement, and advice.

In writing this thesis, I had to draw upon a wide range of prior knowledge I learned long ago, and so I continue to be indebted to my teachers.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

To what extent can a machine model the *human cognitive process* of learning to comprehend sentences of a natural human language, or to associate *meaning* to an English[1] sentence? More specifically, how can a machine learn to assign the appropriate meaning to novel sentences? These are themes this thesis explores.

To explore these themes, a computational system that assigns an appropriate meaning to novel test sentences, and that displays a measure of linguistic *systematicity*, will be proposed in this thesis. The proposed computational system will be specified in detail in chapter 4. The system proposed will incorporate algorithms from the *connectionist* tradition, and will use a *conceptual semantic* analysis of language that is based on *Cognitive Grammar* theory.

Linguistic systematicity is an important cognitive property. It is the property of being able to produce or understand some sentences based on having previously learned how to produce or understand certain others. This cognitive property will be extensively elaborated upon in the rest of this thesis, especially in section 2.4.

To be more specific, the proposed system will be an answer to the research problem of creating a processing architecture, based on artificial neural networks, that:

1. is connectionistic, and can be argued as being oriented towards increasing neurobiological plausibility (this point will be addressed in chapter 5),

2. can be argued as modelling linguistic meaning as imagistic-like *conceptual cognitive structures* for conceptualizing situations — a conception of meaning that has been

[1] We will focus on English, or a fragment of it, as the natural language in question in this project.

argued for by Langacker (2008), Lakoff (1987), and other cognitive linguists (this point will be addressed throughout chapter 4),

3. models the association of the aforementioned cognitive structures with input sentences (this point will be addressed in chapters 4 and 6), and

4. experimentally exhibit at least *partial semantic systematicity*, which also implies *strong systematicity* (These are two kinds of systematicities that humans achieve, and are terms of art to be defined precisely in section 2.4. This point will be addressed in chapter 6).

The research problem described above will be further specified and elaborated upon in chapter 3 as well.

The proposed connectionistic architecture will be capable of modelling the assignment of meaning to novel test sentences on the basis of learning from a training corpus that is *relevantly sparse* (a term that will be explained presently). Training and testing sentences will be generated from simple recursive grammars, and once trained, the architecture will successfully process thousands of sentences containing deeply embedded clauses, therefore experimentally showing the proposed architecture exhibits partial semantic systematicity, and strong systematicity to a significant degree.

Specifically, to test the architecture's abilities to exhibit systematicity and assign meaning to sentences, the architecture will be tested on two distinct grammars. One grammar is based on what was used by Hadley and Hayward (1997), and the other augments what was used by Hadley, Rotaru-Varga, Arnold, and Cardei (2001). In the former, the architecture learns on the basis of 685 unique sentences while it is successfully tested on the basis of 5134 unique sentences. After training, the architecture satisfies criteria for exhibiting the above two systematicity properties.

It must be stressed that "the complexity of the learning process is compounded by the fact (stressed by Chomsky and others) that children learn language under conditions of sparse linguistic input. An important aspect of sparse input is that the set of sentences which a child encounters does not present all words in all syntactically legal positions. Indeed, it seems likely that most words the child encounters are not presented in all legal positions" (Hadley & Cardei, 1999, p. 217). Therefore, the set of sentences the proposed architecture will be trained and tested upon will have to be *relevantly sparse*, meaning that

many words must be restricted from occurring in the training set in what would otherwise be legal syntactic positions, and then be allowed, or even be forced, to appear in those positions in the testing set. Indeed, one of the corpora the proposed system will be experimented upon (namely Corpus 2) will be relevantly sparse with respect to semantic systematicity (the specific restrictions will be defined in section 6.2).

In contrast to other connectionist architectures proposed in previous research (to be surveyed and analyzed in sections 2.3, 2.5, 2.7, and 2.8), the proposed architecture is novel as it commits to an analysis of language meaning on the basis of conceptual semantics, largely as proposed in Cognitive Grammar (Langacker, 2008), and the concept of affirmative stimulus meaning (Quine, 1960) — something none of the surveyed architectures do. The proposed architecture, therefore, demonstrates one possible way of providing a processing model of conceptual semantics, largely as proposed in Cognitive Grammar, that can be implemented within a connectionist framework. How the proposed architecture differs from previous work will be analyzed in detail in chapter 3. The architecture will be argued as oriented towards increasing neurobiological and psychological plausibility as well.

Furthermore, since the experiments in this thesis will provide evidence that the proposed architecture exhibits the aforementioned systematicity properties, and since the proposed architecture is based on connectionist processing principles and conceptual semantic theories, the proposed architecture will be an answer to the challenge, in terms of the variant form of the criticism proposed by Aizawa (1997a, 1997b) that will be discussed in section 2.8.1, for a bona fide scientific explanation of the aforementioned systematicity properties in humans. In contrast, none of the surveyed architectures answer that challenge to the degree the proposed architecture does.

It may seem strange that this thesis will, in part, demonstrate one possible way of providing a processing model of conceptual semantics, largely as proposed in Cognitive Grammar, that can be implemented within a connectionist framework. Stranger still may be that the systematicity properties will be shown through experimentation on artificial languages produced by simple recursive grammars. It is, however, important to stress that quite a few connectionist implementations are of this kind within *cognitive science* (an interdisciplinary field that includes the study of artificial intelligence from the computing science field), as demonstrated by all the works that will be surveyed in sections 2.3 and 2.7, and also the works of, e.g. Stewart and Eliasmith (2009); Smolensky and Legendre (2006); Neumann (2002); Elman (1991); and Elman (1990).

A cursory look at the stated research themes above also reveals a number of concepts that require more careful analysis and explanation. Here is a list of questions that will be considered in this thesis:

1. What is referred to by "human cognitive process?"
2. What qualifies as modelling a human cognitive process?
3. What sort of thing is a meaning?
4. What qualifies as an association of meaning to an English sentence?
5. What is semantic analysis in the conceptual semantics or Cognitive Grammar tradition?
6. What are connectionist architectures?
7. What algorithms from connectionist architectures will be used, and how modified?
8. What is systematicity, why is it important, and can an explanation for its existence in humans be furnished?
9. What qualifies as a sentence in the fragment of English to be studied, as it may be overambitious to make claims regarding the whole of English?

The plan for the rest of this thesis is to first review, in chapter 2, the various fields that form the context of this research; in so doing, we should find answers to some of the questions in the previous list. Next, the precise research problem will be stated in chapter 3 along with further background that shows the specific research question is unanswered by previous research. The approach to answering the research question, including detailed architectural details of the proposed connectionist system, will be provided in chapter 4. Neurobiological plausibility issues will be discussed in chapter 5. Experimental results that show the proposed architecture displays the required systematicity properties will be provided in chapter 6. Concluding discussion will follow in chapter 7.

Chapter 2

Context of Thesis

2.1 Preliminaries

The starting point to understanding the direction of this thesis begins with understanding that the general goal is to explore and deepen, using *machine models* within a naturalistic scientific framework[1], our understanding of human cognition, and more specifically, the *human* ability to comprehend language. By *machine model*, I mean any computational, algorithmic, or *parallel distributed processing* (PDP) network system. It will turn out that we will be interested specifically in PDP networks (also known as *connectionist networks*) due to their supposed neurobiological plausibility, although for now this is unimportant.

As for *cognitive process*, I mean a process of human thinking, i.e. the procedural steps, transitions of states, iterative changes, etc., that occur in *human cognition*, whether the process be conscious, subconscious, or unconscious — to echo Jackendoff (2002), we are interested in studying the *functional mind*, i.e. all mental functions, be they conscious or otherwise. *Human cognition* means the collection of cognitive structures and procedures that operate on those structures; this characterization embraces Thagard's contention that human cognition is composed of "representational structures in the mind and computational procedures that operate on those structures" (Thagard, 2010), but sets aside the issue

[1]By naturalistic scientific framework, I mean a framework of natural laws and forces that are knowable in a scientific and empirical manner within a physicalist universe. I state up front that this thesis assumes a naturalistic scientific framework only to be intellectually honest about what our ontological commitments are. Of course, it is a possibility that the universe is not purely physicalist in nature. It is, however, better to be austere with our initial commitments rather than allow into our ontology, from the very first step, things whose existence is difficult to show openly and empirically.

of whether cognitive structures are intentional and representational (in the philosophical sense). In any case, it is cognition, restricted to those processes dealing with language comprehension, that will be examined in this thesis.

The setting aside of the issue of whether cognitive structures are intentional and representational is motivated by Jackendoff (2002, chapter 2, pps. 19-21, and pps. 279-80), who urged us not to think in terms of intentionality or representations when studying cognitive structures, for the sake of ontological austerity (recall footnote 1). Indeed, Fodor (1999, p. 513) acknowledges that intentional and semantic symbols are *not* compatible with naturalistic scientific theories. It is thus important to note that Quine (1960) assents when "Chisholm counts the semantical terms 'meaning', 'denote', 'synonymous', and the like into the intentional vocabulary" (ibid., p. 220), and further that Quine shows "the baselessness of intentional idioms and the emptiness of a science of intention" (ibid., p. 221) by showing that "the irreducibility of intentional idioms is of a piece with the thesis of indeterminacy of translation" (ibid.). Therefore, in our research into linguistic meaning, we will have to embrace a rehabilitation of the term "meaning" that is more in line with a naturalist scientific and *non-intentional* framework.

In the past two decades, only some work has gone into studying *cognitive* models of language meaning comprehension using connectionist networks. The term "connectionism" has been used variously to describe different algorithmic architectures and philosophical positions. In the next section, we will make this term specific.

2.2 Typical Connectionist Networks

By *connectionist networks*, I mean an information processing architecture or algorithm that has alternatively been called "artificial neural networks", "neural networks", and "parallel distributed processing (PDP) networks", amongst other names. Historically, such networks arose from attempts to model the information processing that occurs in biological systems (Bishop, 2007, p. 226). The obvious appeal of connectionist networks is their *prima facie* similarity with biological neural networks.

The *prima facie* similarity is due to the model being defined as a directed network of interconnected nodes or processing units. Each unit z_j takes input signals sent from units connecting into it, processes the input according to some function, then sends off an output signal to units it connects to. This is *prima facie* similar to how biological neural networks

operate.

As an introduction, I will briefly describe in the following a basic *feedforward neural network* (a directed acyclic graph of neural units), known as a *multi-layer perceptron.* Details of the multi-layer perceptron may be found in Bishop (2007). The description below of the multi-layer perceptrons, absent the specifics of how many layers and nodes to deploy, etc., contains all the typical characteristics of what constitutes a connectionist network.

In a multi-layer perceptron, each processing unit z_j applies a nonlinear *activation function* h onto the *activation* a that the unit receives, where a is a linear combination of the unit's input signals $x_1, \ldots, x_D$, in order to arrive at an output signal (called its *activation value*). That is, if given D units connecting into unit z_j:

$$z_j = h(a_j) \tag{2.1}$$

$$a_j = \sum_{i=1}^{D} w_{ji} x_i + w_{j0} \tag{2.2}$$

We refer to parameter w_{j0} as the *bias*, and w_{ji} as the *connection weight* of the connection from the neural processing unit i to unit j . The nonlinear activation function h is often chosen to be a sigmoidal function, e.g. the logistic sigmoid or the hyperbolic tangent functions, although other functions are often used as well.

Intuitively, the weights represent the strength of the connection between adjacent processing units, used to scale the transmitted numerical signal, while the activation is the sum of the input signals so scaled. Units are arranged into layers, with nodes in layer k receiving signals from nodes in layer $k-1$ and sending signals on to layer $k+1$, if such exists.

In building such a network, a designer must therefore choose the number of layers, the number of nodes in each layer, how nodes in each layer connect to nodes in adjacent layers, and the initial weights and biases of the connections. The initial weights and biases are often set randomly, while nodes are often fully connected to nodes in adjacent layers. The number of layers and nodes to use would be a parameter controlling the complexity of the model.

In addition, the designer must also choose a training algorithm for adjusting the weights and biases during the *supervised training* of the network. *Supervised training* is a kind of training where both the input, and the desired output for each input, are available for use by the training algorithm. In contrast, *unsupervised training* is a kind of training where

the desired output is *not* available during training at all. During supervised training, input information is fed into the network and its resultant output is compared to the desired output; the difference between the two is then used in the algorithm to adjust the weights.

Various training algorithms can be used, many of which are described as *backpropagation* or *error-backpropagation*, referring to how certain information is calculated and propagated between nodes to help them adjust connection weights. For example, the multi-layer perceptron can be trained by applying gradient descent on a sum-of-squares error function (Bishop, 2007, p. 241). There are many other training algorithms. It is important to note that while "backpropagation" might give the impression of information or signals flowing through the network to help improve the network's performance in a kind of biological way, it actually commonly refers to a computational method to efficiently calculate derivatives of the error function — and the "desirability of replacing backpropagation based networks with behaviorally similar architectures which are (at least) closer to biological reality" (Hadley et al., 2001, p. 75) has been noted by Hadley and others (e.g. Hadley et al., 2001), although there is now evidence that some biologically inspired architectures using Hebbian style weight adjustments do exist which can mimic backpropagation (Xie & Seung, 2003).

Given these general characteristics of connectionist networks, we can ask what particular connectionist network architectures are possible. We can basically manipulate the number of layers, number of nodes, how nodes interconnect, which activation function to use, how initial weights and biases are set, and how weights and biases are adjusted during training. With some simple modifications, the network need not be acyclic as well, and *Recurrent Neural Networks* take advantage of this fact.

By *typical connectionist networks*, I will mean to describe networks very similar to ones where each node could be, and often is, fully connected to nodes in adjacent layers if such exist, where a nonlinear activation function is used, where recurrent connections are possibly allowed, and where initial weights and biases of connections are set randomly but possibly subject to certain broad conditions (e.g. that each connection weight be between 0 and 1, or that the connections between two layers satisfy the *echo state* mathematical property). Such networks are often architecturally easy to scale in the sense that the number of nodes per layer and the number of layers may be treated as model parameters that can be adjusted as desired, since the architecture is fairly uniform and not separated into specialized information processing modules individually designed.

Clearly, typical connectionist networks are models that leave no room for the processing

of variables or symbols. Typical connectionist network models are thus "usually described as radically different from so-called classical [symbol processing] AI systems (e.g., Broadbent, 1985; Churchland and Sejnowski, 1989; Clark, 1989; Fodor and Pylyshyn, 1988; Hawthorne, 1989; Hecht-Nielsen, 1990; McClelland, Rumelhart, and Hinton, 1986; Rumelhart, Smolensky, McClelland, and Hinton, 1986; Schneider, 1987; Smolensky, 1988)" (Dawson & Shamanski, 1994, p. 220).

Connectionism is a family of differing positions taken on the question of what extent can various connectionist network architectures model cognition (Memmi, 1990). One such position we may define is what I will refer to as *typical-maximal connectionism*: that typical connectionist networks of a suitable size (i.e. appropriate number of nodes and layers) is "a candidate to cognitive modelling as a whole, and an alternative to classical [symbol processing] AI" (ibid., p. 135) that will "be able to replace classical cognitive models, while offering new qualities" (ibid.). This is essentially the subsumption *maximalist stand* option offered by Memmi, who states, on the relationship between connectionism and classical AI methods, that connectionism would be a better model of cognition, including higher-level processes like language comprehension, and can "do without explicit rules and structured descriptions" (ibid.), but with the stronger condition that it is typical connectionist networks that will prove effective.

Some work has gone into implementing typical connectionist networks to demonstrate their ability to model cognition by way of processing language. In the next section, I will briefly survey three such systems that claim some cognitive or psychological plausibility.

2.3 Typical Connectionist Networks for Language Comprehension

In "Strong Systematicity in Sentence Processing by Simple Recurrent Networks" (Brakel & Frank, 2009), the authors note that "[p]roviding explanations of language comprehension requires models that describe language processing and display strong systematicity" (Brakel & Frank, 2009, p. 1599). They claim even a Simple Recurrent Network trained in standard ways can display strong systematicity, and they built such a network to demonstrate this claim. They based their conclusion on their network's ability to perform well in a next-word prediction task, outperforming the best smoothed Markov models, and conclude that their

network displays strong systematicity without resorting to a symbol processing system that operates over variables.

Strong systematicity is a particular type of *systematicity*, which can be briefly described as the ability of an English speaker to understand "John loves Mary", even if it is a novel sentence to the speaker, given that the speaker already understands "Mary loves John". The *strong* type of systematicity, in brief, specifies that the speaker is able to understand novel sentences containing words, which may appear in embedded clauses, that the speaker has only seen in *other* syntactic positions. A more detailed discussion of systematicity will be presented in section 2.4.

In "Learning Grammatical Structure with Echo State Networks" (Tong, Bickett, Christiansen, & Cottrell, 2007), the authors trained a type of typical connectionist network, called an Echo State Network (which has an architecture very similar to that of Simple Recurrent Networks), to demonstrate their performance in the next-word prediction task. The authors show that an Echo State Network can perform as well as Simple Recurrent Networks, which were demonstrated to be able to handle a weak form of systematicity, without any particularly special training techniques. Based on their results, the authors make some claims to modelling human cognition in that their model gives a partial account for how randomly connected neurons could display complex linguistic behaviours.

In "Connectionist Semantic Systematicity" (Frank, Haselager, & van Rooij, 2009), the authors produced a typical connectionist network (specifically, a recurrent neural network) that they claim models sentence comprehension, and satisfies the important property of systematicity, all without implementing a symbol system. A major innovation in their system is the introduction of a micro-world, so that the connectionist network learns a mapping between sentences and a representation of situations in a micro-world. The authors claim this network learns sentence comprehension in a psychologically plausible manner as it learns the mapping in accordance with structure inherent in the micro-world, arguing that systematicity in language ought to be a result of some equivalent property of the world in which language users inhabit.

In describing these systems, and some of their claims to cognitive plausibility and to sentence comprehension, the property of systematicity was brought up. Tong et al. (2007) actually also made some claims to providing solutions to questions of the nature of linguistic representation. These cognitive claims raise questions of what systematicity is, why it is important, and the nature of mental structures.

2.4 Some Properties of Cognition

Fodor and Pylyshyn (1988) noted that "the ability to produce/understand some sentences is intrinsically connected to the ability to produce/understand certain others" (Fodor & Pylyshyn, 1988, p. 37). That is an instance of linguistic *systematicity*. It is, to use their example, the ability of an English speaker to understand "John loves Mary", even if it is a novel sentence to the speaker, given that the speaker understands "Mary loves John". This property should be read broadly to include such abilities as being able to understand novel combinations of words in embedded clauses even though that combination has never been seen before as a simple sentence or as another sentence's embedded clause. This broad property may be refined and divided into at least three levels — weak, quasi, and strong — as defined by Hadley (1994), although he notes there may be even stronger forms of systematicity in human cognition that cognitive models would need to take into account in modelling language acquisition.

For instance, Hadley (1994) chose "not to include semantic aspects in [his] definition of strong systematicity" (ibid., p. 270) and yet it is clear human language users are, as Hadley later stressed in defining *semantic systematicity*, able to distinguish certain semantic aspects in a way that may perhaps be said to be systematic. A connectionist network may be able to achieve strong systematicity (a syntactic property) and yet fail to achieve *semantic* systematicity, by, e.g. "fail[ing] to distinguish the *meanings* of 'John loves Mary' and 'Mary loves John'" (Hadley, 1994, p. 270).

It seems to me that the "John loves Mary" example above is in fact a simple case of systematicity that a connectionist network may be able to resolve, as the two sentences are dramatically different on the surface (the word order is reversed). A much more difficult case can be seen in attempts to distinguish the meaning between the *de re* (or referential) reading of the sentence "John seeks a unicorn", versus the *de dicto* (or non-referential) reading of the exact same sentence. Given the work, e.g. in Montague (1973) on treating semantics of English through intensional logic and categorial grammar, it seems a symbolic processing system can in fact distinguish between the two different meanings of the sentence "John seeks a unicorn". On the other hand, it is *prima facie* difficult to see how a typical connectionist network could associate two very different meanings to the *exact* same input sentence. This *de re* versus *de dicto* problem, however, is perhaps not a case of a semantic type of systematicity, and is perhaps a case of being able to distinguish two very different

meanings of a semantically ambiguous sentence.

Another example of a semantic ambiguity that human cognition can entertain is demonstrated in the sentence "A negmount[2] has all negmontanic properties" (van Inwagen, 1993, p. 79). The ambiguity is, again, in the meaning of the indefinite article "a" (or "an"), and so the sentence can be read either as "Anything that is a negmount has all of the negmontanic properties", or as "There is a negmount that has all of the negmontanic properties" (van Inwagen, 1993).

From the previous two examples, it is clear that strong systematicity only scratches the surface (e.g. the syntax) of the language abilities of human cognition. As Hadley noted in his example of "John loves Mary" versus "Mary loves John", there is a level of semantic systematicity that needs to be taken into account. Further, with the previous two examples of semantic ambiguity of the *de re*/*de dicto* variety (and there may be others), it is clear that when the meaning of sentences is to be taken into account, the complexity of the problem of modelling human language abilities greatly increases.

To be absolutely clear what we mean by "systematicity", we will use the following definitions (first proposed by Hadley & Hayward, 1997, and Hadley, 1994) for the two levels of systematicity that are most relevant for us in this thesis:

- *Strong Systematicity*: A system is strongly systematic if, after the system has completely learned a training corpus of sentences, it can correctly process a variety of novel simple sentences *without* clausal embeddings, and novel complex sentences *with* clausal embeddings such that the complex sentences will often contain previously learned words in syntactic positions where they *did not appear* in the training corpus (i.e. it is often the case that a word within a novel sentence will *not* have appeared in the training corpus in that very same syntactic position in any of the simple or complex sentences). Furthermore, the size of the training corpus must be small in comparison to the set of novel sentences the system is tested upon.

- *Semantic Systematicity*: A system is *totally* semantically systematic if, after the system has completely learned a training corpus of sentences, it is strongly systematic *and* it assigns the appropriate meanings to all words occurring in *all* novel test sentences that would, or do, demonstrate the strong systematicity of the system. If these

[2] A "negmount" is defined as a necessarily existent golden mountain.

> conditions are met not for *all* novel test sentences, but for a significant fraction of them to a significant degree, then we may say the system is *partially* semantically systematic. Without the "partial" or "total" qualifier, we will assume "semantically systematic" means it is *totally* so, unless the distinction is clear from context[3].

It should be reemphasized that various kinds of semantic ambiguity may not be captured by the levels of systematicity defined above: as Hadley (1994) already noted, there may be even stronger forms of systematicity in human cognition.

Although modelling human language *abilities* is a goal in this thesis, it is important that the model be, to an interesting degree, *psychologically* plausible as well, as my ultimate goal is, generally, to understand human mental and cognitive processes (using machine implemented models within a naturalistic scientific framework). It is thus important that we take a longer moment to think about what the model must achieve beyond creating a *behaviourally adequate*[4] description of human language abilities, and why this is not exactly an exercise in what could otherwise be an application of statistical machine learning algorithms to a problem in natural language processing.

Fodor and Pylyshyn argues that "it seems indubitable that [cognitive capacities] are what we shall call 'systematic'. And... the systematicity of cognition provides... reason for postulating combinatorial structure in mental representation" (1988, p. 37). Thus, it is cognitive structures that are combinatorial, supporting cognitive abilities that are systematic. Our prior focus on systematicity in language comprehension (or production) is, for Fodor and Pylyshyn, simply "[t]he easiest way to understand what the systematicity of cognitive capacities amounts to" (Fodor & Pylyshyn, 1988, p. 37). In fact, "there's every reason to believe that systematicity is a thoroughly pervasive feature of human and infrahuman mentation" (Fodor & Pylyshyn, 1988, p. 37), and so in our modelling of human language abilities, the focus ought to be on studying the structures in the machine that models the cognitive significance of sentences and words being comprehended.

---

[3] Hadley and Hayward (1997) and Hadley (1994) do *not* make a distinction between total and partial semantic systematicity. However, as have been observed in children age 10 or less, even humans are not necessarily *totally* semantically systematic after years of language learning, and thus a "bright-line" distinction of semantic systematicity does not acknowledge the gradation of performance that exist in various systems. This is obviously a distinction that only arises when the study of *unbounded linguistic competence* meets real-world *resource-bounded linguistic performance*.

[4] *Behaviourally adequate* is used here to mean that the output of the machine model is desirable or correct in relation to the given input.

Curiously then, all three typical connectionist network architectures surveyed in section 2.3 make claims to modelling cognition, but their claims rely on their networks' ability to produce desirable output when given certain input. Although their models may be behaviourally adequate — which is debatable, even if their conclusions regarding the performance of their networks are correct — they do not make claims regarding the nature of mental or cognitive states in relation to the machine models of the meaning or cognitive significance of the words and sentences being processed. That is to say, their models make no claims regarding the *psychological reality* of human cognition (although many connectionists may make claims regarding *biological reality* insofar as artificial neural networks are more biologically plausible as models of biological neural networks than, they claim, classical *symbol processing machines*).

2.5 Analysis of Typical Connectionist Networks

The point raised in section 2.4 on the importance of studying the internal cognitive structures and states in the machine model (i.e. artificial neural network) deserve reiterating. Even if we do not believe the claim that human cognition is systematic and combinatorial, studying the structures in the machine that models the cognitive significance of sentences and words should still be the primary concern, *above* that of ensuring the machine model is behaviourally adequate, even as ensuring behavioural adequacy is an important step. Studying thoughts or mental states as modelled by structures in the machine model is important because it is the dynamics of thoughts and mental properties — how they form, how they interact, what their nature is, etc. — that require theorizing and explaining.

Unfortunately, the typical connectionist model ignores the need to analyze cognitive structures in terms of cognitive models of mental states. This is an understandable omission for two reasons. Firstly, "precious little of mind as we ordinarily know it figures in standard connectionist models" (van Gelder, 1993, p. 363), and therefore the empirical evidence from connectionist computational models seems to show that the common-sense "folk" psychological understanding of the mental ought to be entirely eliminated in scientific discussions (the *eliminativist* position), just as scientists no longer speak of caloric and phlogiston. This position has been thoroughly analyzed by Ramsey, Stich, and Garon (1990), and they conclude that if connectionist models that model at the cognitive level are true, then propositional attitudes (e.g. beliefs) simply do not exist (Ramsey et al., 1990,

p. 520).

Secondly, for those who do not wish to go so far as to eliminate talk of the mental altogether, Elman offers evidence that even Simple Recurrent Networks (SRN) have the capacity to represent structural relationships required to process language in a behaviourally adequate manner (Elman, 1991. But see Hadley, 2000, for criticisms against the effectiveness of SRNs). Furthermore, Elman uses the success of typical connectionist models, like SRNs, in processing language adequately (and presumably, also due to their biological plausibility in modelling the brain[5]), to motivate the idea that typical connectionist models can, therefore, also show how *mental* representations come to be constructed (Elman, 1991, p. 221). In particular, the state[6] of the SRN (representing the meaning of the words in a sentence) is a function of the input (the representation of a given word) and the previous state of the SRN's internal representation. Thus each word being processed serves as a cue or signal that causes the SRN state to evolve into, eventually, a representation of the whole sentence being processed, as "[w]ords serve as guideposts which help establish mental states... [and] representations are snapshots of those mental states" (Elman, 1991, p. 221-2). In proposing this view, Elman is clearly relying on identifying the mental representation of a sentence with the connectionist representation, and he made the identification without first ensuring the connectionist architecture is *psychologically* plausible.

Both these positions are seriously problematic, however. In the first case, as van Gelder points out, the argument made by Ramsey et al. depends crucially on the premise that "propositional attitudes [e.g. beliefs] are *functionally discrete*, and *semantically interpretable*, states that play a *causal role* in the production of other propositional attitudes, and ultimately in the production of behavior" (Ramsey et al., 1990, p. 504) — a premise that van Gelder (1993, p. 365) argues to be false, and has been so argued by others, including Dennett (1987a), for example. Arguing here why this premise is false would certainly take us much further afield than would be appropriate, but the fact that strong arguments exist to support that it may very well be false certainly shows us how important it is to gather further empirical analysis of the relationship between the representations or structures in connectionist models versus those found in psychological models.

---

[5]Recall that the biological plausibility of connectionist models is a major "selling" point of this kind of modelling to begin with.

[6]Elman presents evidence that the connectionist processes involved must be using internal representations that are *distributed* in nature. Distributed representations will be discussed in this section below.

In the second case, Elman's identification of mentality with connectionist representations is problematic because it seems it would leave us with having to identify mental states with neurobiological states when connectionist models become sufficiently neurobiologically realistic (through active and continuing research) as to be identifiable with neurobiological models of the brain. That is, we are left with having to subscribe to a *psycho-neural identity* theory of the mind, which is a class of theories that are strongly undermined by a number of philosophical arguments aimed at certain key foundational issues (e.g. see Kim, 2005, pps. 106-13). Even if we ignore the foundational arguments for the time being, the simple identification of mental states with brain states, modelled by connectionist models, without at the same time ensuring that the connectionist models are psychologically realistic leaves us in an odd position: unlike the first case, we are here committed to the existence and realism of mentality, and yet we have no plausible model of it, if all we have is, by hypothesis, a biologically realistic model of the brain that has not been shown to be psychologically realistic as well (a simple appeal to the identity between brain and mental states to solve this oddity is, of course, question begging). Clearly, behavioural adequacy of connectionist models is insufficient in cognitive modelling if we wish to understand both the psychology *and* the neurobiology.

Fortunately, the hypothesis that cognition "can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures... is general enough to encompass... connectionist theories" (Thagard, 2010), and of course the same hypothesis with "representational" edited out (so that intentionality is not referenced) is even more general. In fact, whether Ramsey et al. are correct that beliefs do not exist, or van Gelder and Dennett are right in arguing the opposite view, what is urgently necessary for informing the debate is more analysis of the relationship between connectionist and psychological models of cognition: after all, there is every reason to take into account advances in connectionist research of cognition when deciding on the ontological nature of beliefs, as van Gelder (1993, p. 365) points out. A similar conclusion can be drawn, of course, from the inadequacy of connectionist models that are behaviourally adequate but developed in isolation from psychological theories, as demonstrated through our analysis of Elman's claims described above.

Therefore, we must remember that it is the cognitive structures and computational procedures, and their relationship with more psychologically plausible theories of cognition, that are of interest in addition to seeking greater behavioural adequacy, and that are ends towards

which all the typical connectionist network architectures surveyed in section 2.3 made no claims.

Making no claims is, of course, not reason enough to reject all typical connectionist networks (or even just those surveyed in section 2.3) as inadequate for modelling cognition. It seems, however, that there is significant evidence against the typical-maximalist connectionism position that typical connectionist networks are adequate for modelling cognition as a whole. For example, Hadley (2000) analyzed three different classes of artificial neural networks (including SRNs like the one used by Elman, as discussed above), two of which are of the typical connectionist network type, and concludes that they are inadequate to modelling, especially, higher cognitive functions due to their reliance on infinite precision, methods to acquiring connection weights that are implausible for high-level cognitive functions, etc.

Fodor and Pylyshyn (1988) offered a more general analysis of connectionism as a model for cognition, critiquing some of the usual reasons for preferring connectionist networks over *Classical architectures*, such as resistance to noise and damage. More importantly, they conclude that a key aspect of the dispute is in whether cognitive models ought to include mental representations that possess a combinatorial structure (syntax and semantics), and whether mental operations and processes are sensitive to that structure. *Classical* models of cognition commit to a model with mental representations that have explicit combinatorial structure, and with mental operations that are sensitive to these structures; typical-maximal connectionist models deny having either properties.

Instead, typical-maximal connectionist models, like that of Elman (1991), rely on analyzing mental representations in terms of the encoding scheme employed for the input, output, and other layers of neural units. The encoding is either *localist* or *distributed* in nature. "In localist representations, each input node corresponds to a specific word or concept, and only one input node is activated at a given time. In distributed representations, the inputs are encoded by sets of nodes, with each input node corresponding to a feature; an individual entity corresponds to a set of simultaneously activated features that typically represent subcomponents such as phonological or semantic units" (Marcus, 1998, p. 245). Most typical-maximal connectionism research uses either local or distributed encodings for the task of mapping from an input vector to an output vector in a supervised training style of learning (Marcus, 1998). For example, the SRN used by Elman (1991), as already discussed above, uses a distributed encoding, and Elman argues that the use of distributed

representations is key to achieving behavioural adequacy.

Note however that my arguments above regarding the need for greater research into the relationship between connectionist and psychological models of cognition does *not* depend on the implementational detail of the encoding scheme employed in the connectionist architecture. For this reason, discussion of localist versus distributed encoding is insufficient, even if important. Be they localist, distributed, or combinatorial, the cognitive structures found in connectionist networks, or found during the processing of a connectionist network, must be examined more closely. For all the above reasons, at least in terms of higher cognitive functions that allow humans the ability of process symbols and language, *modelling the human cognitive process* ought to be taken to mean providing a model of the formation and nature of cognitive structure.

It seems this is an area lacking in most of the research into typical connectionist networks for cognitive modelling. As an example, although the importance of mental representations is directly mentioned by Frank et al. (2009), they too did not analyze the representations that may have been learned within the Simple Recurrent Network they employed. They instead hung the claim that their network has no combinatorial structured representations upon the fact their network's *output* has no such structure (but surely the input and output having no combinatorial structure does not in itself prevent such structure from arising *within* the network). There has been some research into the nature of representations in typical connectionist networks, though according to Fodor and Pylyshyn, none have adequately answered their challenge of explaining systematicity and structure sensitive processing (Aydede, 1997).

Still, it is important that we explore these issues specifically through connectionist models, for they have, in the past, been seen as possibly more biologically plausible as models of information processing in biological systems (Bishop, 2007, p. 226). They have also been held as better models for "a more plausibly mammalian cognitive architecture" (Dennett, 1991, p. 28) overall, and so they should not be so quickly dismissed as models for human cognitive processes since the higher cognitive functions in humans are likely "a recent add-on" (ibid.).

Since we are interested in higher cognitive functions, and in particular sentence comprehension, we may pause to wonder what it is that the cognitive structures are, or at least what function they serve in sentence comprehension. We may hypothesize that within the realm of sentence comprehension, the cognitive structures are the meaning of sentences or

sentence-parts, since we are committed to studying cognitive significance or thoughts as modelled by cognitive structures.

Given, as Fodor and Pylyshyn (1988) noted, that humans behave with respect to language in a systematic and productive manner, we might conclude that the mental representations ought to possess a combinatorial structure, and that mental operations over them are sensitive to these structures — at least that is what the Classical models of cognition commit to — whereas typical-maximal connectionist models may explain the systematic and productive linguistic behaviour as being implemented through some kind of processing of distributed representations in a connectionist architecture. In characterizing the debate this way, we may of course be heading for a false dichotomy. So before proceeding further, we may rightly ask just what sort of thing is a meaning, and whether the typical connectionist networks proposed as models can adequately account for it. These are important questions since, as argued above, any satisfactory model must seek psychological plausibility while it is still important that we explore these issues specifically through connectionist models.

2.6 Language and Meaning

What sort of thing is sentential meaning has been studied in depth both in the philosophy of language and in linguistics. Of the many proposals, it seems there are four that are of special interest, namely semantic marker assignment, model-theoretic semantics, truth-theoretic semantics, and conceptual semantics.

Some linguists have conceived of "semantic interpretation as the assignment to sentences and their constituents of compounds of 'semantic markers' or the like" (Lewis, 1970, p. 18) where "[s]emantic markers are *symbols*: items in the vocabulary of an artificial language we may call *Semantic Markerese*" (ibid.). Thus an association of meaning to a sentence consists of associating a formula in the *Semantic Markerese* to the English sentence, perhaps translating the English to the Semantic Markerese via a mechanism akin to the *generative grammar* phrase marker trees (see J. J. Katz & Postal, 1978, for instance).

There seems to be two general problems with any theory of meaning that takes meaning to be semantic markers. First, as Lewis noted, any such translation from English to any Semantic Markerese is tantamount to translating from an object language to an auxiliary language, and therefore does not constitute a proper theory of meaning, since "we can

know the Markerese translation of an English sentence without knowing the first thing about the meaning of the English sentence: namely, the conditions under which it would be true... Translation into Latin might serve as well" (Lewis, 1970, p. 18) (see Vermazen, 1967, for similar criticisms and in more detail).

The second problem is related to the first, being that some kind of Semantic Markerese seems to be assumed in many typical connectionist network models of language comprehension. This is most easily seen in the local representations of words in any connectionist network: if the meaning of an input sentence is taken as the activation of certain neural units, e.g. the units labelled i, j, and k, then the network has translated the input English sentence into the Markerese sentence ijk, or if we choose a different "dialect", into the Markerese symbol x where x is "spelled" $(0, \ldots, x_i, 0, \ldots, x_j, 0, \ldots, x_k, 0, \ldots)$ with 1 in place of x_i, x_j, and x_k. The same criticism applies as before in that we could know the translation of an English sentence into this Markerese without knowing the first thing about the meaning of the English sentence.

Insofar as we are concerned with the translation of one language to another, there is nothing wrong with using connectionist networks and assume semantics as an association with Semantic Markerese in order to tackle this important task as an application in natural language processing. If we are to model cognition and meaning, however, it seems we must examine other proposals of what meaning is.

A different proposal is the *model-theoretic semantics* in the tradition of Frege (see Frege, 2000; Lewis, 1970; Partee, 1973; Montague, 1973). The full proposal in the form Montague and others arrived at employs the use of a *categorial grammar* with a *transformational* component, *Carnapian intensions*, and *intensional logic*. At its simplest, we may understand the proposal as being defined by a commitment that a large part of the meaning of a word is a function mapping certain features of the entire world (organized into an *index*) to the extension of that word; and the meaning of a sentence is a function of the functions that are the (more or less) meaning of the words in the sentence.

The softness of the above explanation is due to the technicality of the proposal and a more complete spelling out of the proposal would be required if we were to use it as our understanding of what a meaning is in order to implement a connectionist network. To add in some detail though, note that these functions are not the whole story to meaning, for meaning will necessarily encompass more than the intensions of words, and include *pragmatics*, meaning from implications (Gricean *implicatures*), etc. So in this proposal,

"meaning" is restricted to being identified with phrase marker trees, that have assigned to its nodes intensions appropriate to the sentence being represented.

Some of the problems in employing this proposal within a connectionist network begin with trying to represent an *index*, which is a n-tuple "of the various items other than meaning that may enter into determining extensions" (Lewis, 1970, p. 24) including possible-worlds, time, place, speaker, audience, indicated-objects, etc. We then have to represent functions (the intensions) in such a way that allows for function composition (and in fact, some of the functions turn out to be higher-order functions). Further, these functions must be "held" within the connectionist network in such a way that they are identifiable and not merely used (i.e. the neural network is not being used to learn a function to process individuals in the input, rather the network must learn a function to process functions. In programming parlance, functions must be treated as first-class objects).

The last point is important: employing model-theoretic semantics, the network will have to learn a function that can process functions, and *not* simply learn a function that processes individuals. The difference is between having a theory of meaning and not having one at all. Frege's initial proposal called for the use of functions to calculate the meaning of sentences from the meaning of words rather than explicitly identifying the intensions of words and sentences as functions that do not have meaning as value. The result was criticized, for example by Davidson, as one may "[a]sk, for example, for the meaning of 'Theaetetus flies'. A Fregean answer might go something like this: given the meaning of 'Theaetetus' as argument, the meaning of 'flies' yields the meaning of 'Theaetetus flies' as value. The vacuity of this answer is obvious" (Davidson, 1967, p. 304). The resolution of the problem is seen by identifying the "meaning" (by which I mean the intension) of "Theaetetus" as a certain function t (mapping indices to things), and identify the intension of "flies" with f, so that the "meaning" of the compound is $(f(t))(i)$ where $f(t)$ is a function that takes as argument i, which is the appropriate index (containing relevant features of the world), and has as value either Truth or Falsity. Armed with the function as the intension of the sentence and the appropriate index, we may calculate whether the sentence is true or false, showing that this construction of meaning is not vacuous.

The difficulty of learning higher order functions in a connectionist network aside, it is not clear that a cognitive agent could even form a proper index with which to understand a sentence, or even acquire the relevant higher order functions. Indeed, the formal intensions required to be calculated, in fact, are not even usually Turing computable. Specifically,

for instance, due to "the computational intractability of the infinite constructions required by... possible world semantics" (Jackson, 1996, p. 135) (See also Hadley, 1989, p. 134). Further, the model-theoretic tradition (and indeed, most traditions tracing back to Frege. See Jackendoff, 2002, pps. 296-300), conceive of language as an abstract entity in the world, that makes reference to objects and states of affairs in the world. Language users as cognitive agents come into the picture only as their minds are needed to "grasp" this abstract entity. It is not spelled out just how a mind should, in a naturalistic manner, be able to "grasp" the entirety, or a sufficient portion, of the possible worlds model required to understand and use language, given that the cognitive agent has access only to the one actual world.

Even if the study of language as an abstract entity "out there" outside the mind is a worthy pursuit (and it is, as demonstrated by the advances in statistical natural language processing, machine translation of natural language, etc.), if we are to understand sentence meaning as "grasped" by language users, we will have to make a better attempt at studying meaning on its own terms. That is, in terms that are cognitively plausible.

An alternative proposal from Davidson proposes, in summary, that we conceive of the meaning of a sentence as the set of *T-sentences* relevant to understanding the truth condition of the sentence in question. This is called a *truth-theoretic* model of meaning. *T-sentences* are the sort defined by Tarski as the meaning of the "is true" predicate. Thus, e.g. "'Snow is white' is true if and only if snow is white" would be one of the T-sentences needed to understand the sentence "Snow is white" (Davidson, 1967).

From a connectionist point of view, this is actually an interesting proposal, as features of the world identifiable by certain nodes to indicate the whiteness of snow in the world can be associated, through using an artificial neural network (ANN), with the sentence "Snow is white", effectively having the ANN learn the T-sentences directly. It seems many of the networks that employ a Semantic Markerese theory of meaning could be re-purposed to explore this truth-theoretic construction so long as a convincing cognitive modelling of the world, less language, could be constructed.

One immediate difficulty in employing the truth-theoretic proposal within a connectionist modelling of cognition is, of course, in creating a convincing cognitive model of at least some fragment of the world, less language. It would not be enough, as Frank, et al., did to simply encode or map situations in a micro-world described by a Semantic Markerese to situation vectors, and then claim, as they did, that the mapping from Markerese to

vectors "is not intended to simulate the psychological process of developing event representations" (Frank et al., 2009, p. 12). Doing so seems to beg the question insofar as we are studying meaning as *cognitive structures*.

Another apparent difficulty is that, parallel to the model-theoretic case, truth conditions are not in general effectively computable. (Similar concerns in regards to Procedural Semantics are persuasively critiqued by Fodor, 1978, and attempts to solve it are, for instance, presented by Hadley, 1989, and Wood, 1981). This is not, however, a major difficulty as Davidson argues that:

> Even if we hold there is some important sense in which moral or evaluative sentences do not have a truth value (for example, because they cannot be "verified"), we ought not to boggle at "'Bardot is good' is true if and only if Bardot is good"; in a theory of truth, this consequence should follow with the rest, keeping track, as must be done, of the semantic location of such sentences in the language as a whole — of their relation to generalizations, their role in such compound sentences as "Bardot is good and Bardot is foolish", and so on. (Davidson, 1967, p. 316-7).

We need not boggle because Davidson proposes that users learn the meanings of a language by following a principle of charity, that of assuming sentences spoken by native speakers are true, and then by building up a dictionary of T-sentences, the user can work out a theory of truth that would allow the user to deduce the truth-conditions of other sentences[7].

A more substantive difficulty does arise, however, from how the truth-theoretic proposal couches the study of meaning in the study of deductive truth, when in fact, statements of meaning are *stronger* than statements of truth. Consider, for example, that the sentence schema "'S' means that p", combined with the a priori analytic schema "If 'S' means that p, then 'S' is true if and only if p", entails the standard T-sentence schema "'S' is true if and only if p". The deductive derivation in the opposite direction is simply not possible (Soames, 1992, p. 17). In order to justify what we may now see as an induction, Davidson proposes that meaning is in fact *holistic*: i.e. the meaning of a sentence depends on the semantically significant parts of the sentence, and the meaning of the parts depends on their systematic

[7]It may now be clear how "connectionist friendly" Davidson's proposal could be, since the dictionary of T-sentences may be interpreted as a data-set useful for supervised training of an ANN.

contribution to the meaning of *all* sentences in the language in which they occur (Davidson, 1967, p. 308).

More difficulties, however, abound for this proposal, including how pieces of language can succeed in referring to actual objects or events when absent a cognitive agent (language is, in this proposal, still thought to be an abstract entity external to minds); and how to "filter" out non-translational T-sentences that crop up due to the identification of a theory of truth as a theory of meaning, as may happen if a speaker mistakes all "'S' is true iff p" T-schema with the usually non-translational schema "'S' is true iff p and arithmetic is incomplete".

All the difficulties mentioned above that arise from analyzing the relationship of meaning and truth in a truth-theoretic theory of meaning may seem to take us pretty far afield, but I mention them in some detail above to emphasize three points. First, meaning is holistic in a theory of meaning that deduces the meaning of words from the meaning of a set of known sentences such that the meaning of novel sentences are deduced from the meaning of the words used. Second, that a theory of meaning that depends on a theory of truth raises philosophically foundational difficulties that are quite enormous, and makes Davidson's proposal mostly untenable for use in naturalistic cognitive science. Third and finally, all these difficulties should serve as motivation for asking an important question: why does truth need to figure into the translation at all? In fact, the same may be asked of model-theoretic theories that posit the meaning of sentences as either Truth or a function that maps to Truth — why map to Truth or Falsity? This question is all the more important since my goal is to better understand the human cognitive ability of comprehending language, not to entertain theories of absolute truth.

Although taking the reference of declarative sentences as a truth value is customary in formal semantics, going all the way back to Plato (Jackendoff, 2002, p. 328)[8], it is actually neither necessary nor is it fruitful for a thoroughly *cognitive* understanding of meaning, and "[t]heorists' concentration on truth value... blinds us to the full vivid range of possibility" (Jackendoff, 2002, p. 328). In fact, *conceptual semantics* is an alternative that has been argued as being more fruitful for studying meaning as a *cognitive* phenomenon, in which "the intended reference of a declarative sentence is a situation (an event or a

[8]In fact, Frege apparently took truth values as existing abstract entities (i.e. The True, and The False) that declarative sentences refer to (B. Linsky, personal communication, 26 January 2005).

state of affairs)" (Jackendoff, 2002, p. 326). A situation[9] is an observable, or in principle observable, phenomenon, and so it may appear as though we are starting to move from talk of linguistic meaning to talk of pragmatic meaning. Under traditional categorizations, this may indeed be true. As we are studying meaning as concepts or items "held" in the mind or cognition, rather than as formal mathematical objects or socially shared creations, the separation of linguistic meaning from conceptualizations is not viable, and "we must consider the domain of linguistic semantics to be continuous with human conceptualization as a whole" (Jackendoff, 2002, p. 282), so that "semantics and pragmatics form a gradation... with no precise boundary between the two" (Langacker, 2008, p. 40). A brief argument for this continuity follows, but a full discussion would take us too far afield (see Jackendoff, 2002, pps. 281-93; and Langacker, 2008, pps. 36-43).

Here is a sketch of an argument that we should *study* linguistic and pragmatic meanings as parts of a continuous domain, but note this is *not* an argument that they are *in fact* a continuous domain (i.e. this is an argument in support of our research *methodology*). Suppose to the contrary that linguistic and pragmatic meanings are distinct. Further, suppose they are both concepts "held" within the mind (as opposed to being something external to the mind). Then there must exist two distinct types of cognitive structures, each specific to one of the two distinct kinds of meanings. How can we *empirically* establish the existence of these two kinds of cognitive structures? One way is to build models, but it would be question begging to build cognitive models of language learning that only model linguistic meaning, and not pragmatic meaning, and then to conclude that they are therefore separate[10]. Thus, on methodological grounds, it would be more fruitful to engage in a research program that builds models that allows for the inclusion of *both* kinds of meanings *without* presuming a special distinction between them. To conclude this sketch, note that Elman (1991) has in fact provided evidence that meaning is contextual and pragmatic, and that a distinct linguistic meaning might not be necessary for higher levels of abstraction[11] (he even

[9]Unless a precise distinction is required between a situation as a whole versus a fragment of it, the term "situation" will be used for both cases without specifying the distinction. This allows us to conveniently say that, e.g. "The cat is on the mat" refers to a situation, without having to also mention that it is under assumptions of standard viewing arrangements, or that it is dependent or independent of certain standard conditions, etc.

[10]Of course, that is unless such a model is *completely* behaviourally adequate and biologically realistic, but, to date, no such model exist. If such a model did exist, however, then we would be faced with the prospect of having to eliminate one of the two types of meanings on empirical grounds via Occam's razor.

[11]Elman argues that connectionist models "emphasize the importance of context and the interaction

created his model without presupposing two distinct types of meanings). Elman's evidence is based on a simplistic model, with conclusions that are questionable: and that precisely demonstrates the importance of the preceding methodological argument's conclusion.

Furthermore to understanding how conceptual semantics can have the intended reference of a declarative sentence be a situation without invoking non-physicalist principles, notice that the situation is only *intended* as the reference. As already discussed at the end of section 2.1, reference (as denotation) is *incompatible* with naturalistic science[12] due to it being a part of the intentional idiom. The key fundamental insight to conceptual semantics as posited within Cognitive Grammar theory is that the basic components of grammar are pairings[13] between semantic structures (which Langacker also calls "meaning" and "conceptualizations") and phonological structures (which Langacker also calls "phonological shape") (Langacker, 2008, p. 5), such that the semantic structures are sometimes physically caused by the situation. It may seem, at first, like Langacker is playing a game of relabeling in "[a]dmitting that meaning resides in conceptualization" (ibid., p. 31), since linguistic meaning is traditionally (as seen in the above survey of different theories of meaning) viewed as "transcendent, existing independently of minds and human endeavor" (ibid., p. 28), and residing in abstract entities external to minds (i.e. they are platonic); or is viewed as residing in sets of objective truth conditions about the world irrespective of how the world might be humanly seen and experienced (so in the traditionalist views, words can denote real things and sentences can refer to real situations).

Finding meaning as something completely within the mind, however, is just psychologizing the study of meaning into what Jackendoff calls a *mentalistic enterprise* (Jackendoff, 2002, p. 267). It is apparently in line with what Fodor had in mind in advancing a conceptualist semantics that made use of a Language of Thought (LoT) to mediate between language and the world, through the mind of a language user (Jackendoff, 2002, p. 300). It

---

of form with meaning... As [his SRN model] demonstrates, these characteristics lead quite naturally to generalizations at a high level of abstraction where appropriate, but the behavior remains ever-rooted in [distributed] representations which are contextually grounded" (Elman, 1991, p. 221).

[12]So construed, some philosophers argue that naturalistic science can only address a subset of "ultimate reality", and some further believe such a stance to be tendentious and false. Note the idea that naturalistic science can only address a subset of "ultimate reality" is not necessarily a bad or indefensible outcome (interested readers may see, e.g. Nagel, 2002; Midgley, 1994).

[13]Langacker (2008, p. 5) calls a pairing of semantic and phonological structures a "symbol", but I will refrain from using his terminology of "symbol" in this thesis in order to reduce confusion in the name clash that would occur, since others have used "symbol" to mean (approximately) a word having as meaning something in the external world (as opposed to having as meaning a structure that is purely within cognition).

seems to also be in line with what Hadley called the nature of semantic competence (Hadley, 1989, p. 118). Ultimately, cognitive research is interested in what goes on in people's heads, and not in platonic ideals that somehow float above them.

Although Fodor's Language of Thought are *mental* representations, they are representations of real things or situations. The fact that the LoT is supposed to have a formal syntax and use mental *representations* means that it has to have a semantics that interprets or maps its expressions into the real world (that is what makes representations *representational*) (Jackendoff, 2002, pps. 278-9). The LoT is thus open to the Semantic Markerese criticism discussed earlier. Furthermore, as Jackendoff argues, being representational and having an interpretive semantics inherits into the LoT all the problems and criticisms that the realist or model-theoretic proposals had with truth and reference, but then also adds into that mix the problems from having to contend with intentionality (given the problems of intentional idioms as discussed at the end of section 2.1, it does not help that "Fodor insists that LoT is *intentional*: it is *about* something" Jackendoff, 2002, p. 279), and thus of consciousness[14]! Therefore, in this thesis, we will instead lean on the conceptual semantics understanding of meaning, as analyzed through Langacker's Cognitive Grammar theory (Langacker, 2008), and through Lakoff's analysis of various elements in conceptual semantics (Lakoff, 1987); that is to say, linguistic meaning resides in conceptualization.

The difference from Fodor is that, in Cognitive Grammar, the semantic side of language (i.e. what Langacker calls "semantic structures" or "conceptualizations", which is what is paired with what Langacker calls "phonological structures" or "phonological shapes") is *not* in itself a formal language with its own formal syntax and interpretive, and indeed intentional, semantics that requires naturalizing into non-intentional and scientific terms. Rather, *conceptualizations* interface *physically* with *perceptions* of the world ("*perceptions*" here defined as signals generated by the vibration of the eardrum, retinal signals from *ocular irradiation*, which is the pattern of photons hitting the eyes, and other signals from other sensory organs, neurobiologically processed in the brain). "Ultimately, conceptualization resides in cognitive processing. Having a certain mental experience resides in the

[14]Consciousness may be required for a complete model of cognition (Hadley, 2009), but for the sake of metaphysical and ontological austerity, let us attempt to advance as far as possible our understanding within a naturalistic scientific framework and a physicalist ontology, before making an appeal to consciousness.

occurrence of a certain kind of neurological activity[15]. Conceptualization can thus be approached from either a *phenomenological* or a *processing* standpoint: we can attempt to characterize either our mental experience per se or the processing activity that constitutes it. Cognitive semantics[16] has focused on the former... As for processing, it can be studied at different levels (both functional and neurological)" (Langacker, 2008, p. 31), and this thesis is committed to understanding the processing aspect using a computational model that is hopefully oriented towards increasing neurobiological and psychological plausibility (the processing aspect of conceptualizations will be further characterized in section 4.1).

It may be argued by some that the semantic structures proposed in conceptual semantics is just a kind of Fodorian Language of Thought (LoT), and it is an exercise in relabelling to call it meaning. However, the claim here is that these semantic structures or conceptualizations, which as described previously are physically caused from the neurobiological processing of perceptions, are not intentional or representational of external reality in any special philosophical sense. As Fodor and Pylyshyn (1988) made it quite clear, Classical theories postulate a Language of Thought or mental representation where "the semantic content of a (molecular) [mental] representation is a function of the semantic contents of its syntactic parts, together with its constituent structure" (ibid., p. 8). Thus, in Classical theories, the "expressions in [Fodorian] LoT are *mental representations*, and they *represent* something: entities in the world. Put differently, Fodor insists that LoT is *intentional*" (Jackendoff, 2002, p. 279), it has "semantic properties (i.e. are meaningful, can be interpreted as being about something)" (van Gelder, 1990, p. 366), and it has *semantic content*.

Notice how drastically different the Classical view is from what is proposed by the conceptualist semantics of Jackendoff (2002), Langacker (2008), et al. In conceptualist semantics, an English sentence utterance generate phonological cognitive structure that gets associated with a conceptualization cognitive structure. Conceptualization cognitive structure does not have meaning, does not have semantics, does not have semantic content, and is not intentional; however, conceptualization cognitive structure *is* the semantics, or

[15] Although conceptualization can be approached from either phenomenological and processing standpoints, we need not subscribe to a naive theory that *identifies* mental states with neurobiological states. For one treatment of this issue that "saves" folk psychology within a physicalist ontology, given the irreducibility of intentionality to physicalist terms, see Dennett (1987b).

[16] *Cognitive semantics* is the conceptual semantics as proposed in Cognitive Grammar theory. In this thesis, we make use of only a small subset of Cognitive semantics.

put differently, the *cognitive significance*, for English sentences *as utterances* that generate certain phonological cognitive structures[17].

It may be objected that symbolic Markerese, from Classical architectures, also *is* the semantics for language, but the real issue here is that symbols in Markerese requires its own *separate* (secondary and auxiliary) interpretive formal semantics to enable the Markerese symbols to *represent* and be *about* things in the external world. Conceptualization cognitive structure does not have such a separate intentional semantics *at all* (see also section 4.1).

Having said that, R. F. Hadley (personal communication, 29 May 2011) suggests "that many researchers, including Harnad and [himself], hold that a large part of semantic theory must be devoted to explaining how semantic grounding of a sizable subset of our internal representations must occur in order for semantic representations to exist within the brain at all. Such semantic grounding must include how some of the internal representations *become* representations via their relationship to possible *denotata* in the external world. Such relationships have been variously analyzed in terms of standardized causal processes (with counterfactuals being involved), and these processes might involve quasi-computational procedures, and/or neurological training, etc." To continue to pursue the issue of *semantic* or *symbol grounding*[18] would take us much too far afield in this thesis, and so I will not discuss these issues much more except to note the following: Conceptualization in Cognitive Grammar's conceptual semantics is neurobiologically generated from stimulus received from external physical objects and events (e.g. through ocular irradiation), and so conceptualization as envisioned by "cognitive linguists is noninsular, being grounded in perception and bodily experience" (Langacker, 2008, p. 28). Much more details on conceptualizations will be presented in section 4.1, but as only a very small subset of Cognitive Grammar is required for the purpose of this thesis (wherein an *incomplete* cognitive model of language learning is being built), the issue of grounding will not be pursued much further (but see Langacker, 2008, chapter 9, for a detailed analysis of grounding in Cognitive Grammar).

[17]This echos the analysis by Jackendoff, who argued that: "Semantic/conceptual structure does not *have* a semantics, it *is* the semantics for language" (Jackendoff, 2002, p. 278-9).

[18]It should be noted that whether or not conceptualizations or LoT symbols are grounded is not the same issue as whether they have intentionality. Discussion of this point is beyond the scope of this thesis and will not be pursued further.

2.7 Atypical Connectionist Networks for Language Comprehension

So far our attention has been directed towards typical connectionist networks, which, as described in section 2.2, have architectures that are usually fairly uniform and not modularized in any significant manner (beyond the typical separation into layers of nodes). We now turn our attention to connectionist network architectures that may be described as *atypical*, some of which are even *hybrid* (Hadley, 1999) in that they contain modules that function autonomously from each other.

The particular properties that make a network *hybrid* are that at some higher abstract level, the network can be described as having a classical architecture, where information flows between separate modular sub-networks, while at a lower-level some of those modules may function autonomously, even if their functioning may causally impinge on the functioning of other modules, and while some of those modules may be connectionistic, not all of them have to be so. At the higher level, it is clear that the network is not of a typical connectionistic design, but it still cannot be said that it is purely classical in nature, as "emergent properties not found in purely classical architectures" (Hadley, 1999, p. 212) may be found due to the effects of having modules that function in a typical connectionistic manner.

Atypical connectionist architectures share, in common with hybrid ones, the property that the whole is separated into modules, where each module may be of a completely different design. However, some atypical connectionist architectures are not hybrid as their modules do not function autonomously, and often times, none of the modules employ classical symbol processing architectures at all. Furthermore, as opposed to typical architectures, atypical architectures do not easily scale in architectural size, in the sense that typical architectures can easily be adjusted in the number of layers, number of nodes per layer, etc.

For instance, *Deep Belief Nets* are a kind of neural network design that can be described as a stack of either restricted Boltzmann machine, or auto-encoding, learning modules. Weights of the lower layers are learned, then fixed, before the next layer of connection weights are learned, and so on up the stack for as many layers as desired (Bengio, 2007, pps. 35-7). The regularity of the design, allowing for an indefinite number of learning modules stacked on top of one another, and the uniformity of how each function and impinge on the functioning of adjacent layers, make it a *typical* connectionist architecture.

Several atypical connectionist networks have been studied before that show surprising properties in language processing. A survey of four notable ones are below.

As a first example, Hadley et al. designed a "Hebbian-inspired, competitive network... which learns to predict the typical semantic features of denoting terms in simple and moderately complex sentences" (Hadley et al., 2001, p. 73) that can be described as atypical. It works in an *error*-unsupervised[19] fashion, and it achieves a strong form of systematicity. The part of the design that is most interesting is the creation of two modules or layers of nodes whose purpose is to "serve as memory copies" (Hadley et al., 2001, p. 77) of the first hidden layer. Although recurrent networks also provide a form a memory, the finite nature of the memory provided by having only two modules may perhaps prompt some to argue that it is an approximation of a human short-term memory function.

As a second example, Hsiao (2002) designed an atypical connectionist network that employs two sub-networks, each of which is a version of the network previously designed by Hadley et al. (2001). By using different parameter settings, the two sub-networks can "discern coarse-grained and fine-grained categories respectively" (Hsiao, 2002, p. iii), allowing one sub-network to have "a greater capacity for recognizing the syntactic structure of the preceding words, while the other will have a greater capacity for recognizing the semantic structure" (ibid.). Furthermore, an interesting "mechanism to switch attention between the predictions from the two sub-networks, in order to make the global network more closely approximate human behavior" (ibid.) is included. One might describe such a module, which in some sense watches over the two sub-networks, to be reminiscent of a meta-cognitive ability that has been identified as important in student learning.

As a third and more notable example, Hadley and Cardei (1999) designed a *hybrid* connectionist network with four modules where the input-layer feeds into a Kohonen Map and into a Concept Layer; the Kohonen Map also feeds into the Concept Layer; the Concept Layer feeds into a Feature Layer that also feeds into the Kohonen Map; and crucially, the network also employs a classical message passing algorithm between neural nodes to facilitate the process of binding concept nodes to thematic role nodes using *conjunctive binding nodes* — nodes that fire cyclically after receiving sufficient input from nodes that are to be bound together. Hadley and Cardei showed that a network of such an architecture was

[19]Describing a training algorithm as *error*-unsupervised training is *not* the same as describing it as unsupervised training. An error-unsupervised training algorithm, e.g. Hebbian training, can be used in *both* supervised and unsupervised training situations.

able, "on the basis of (relevantly) sparse input, to assign meaning interpretations to *novel* test sentences in both active and passive voice" (Hadley & Cardei, 1999, p. 217) through *error*-unsupervised training, and was shown to display strong semantic systematicity.

Also of note from Hadley and Cardei (1999) is their careful study of the representations in the output Concept Layer that became associated with the input sentences. The Concept Layer is designed in such a way as to allow the generation of representations with appropriate concepts (e.g. dog, cat), implemented as concept nodes, be bound by binding nodes to appropriate thematic roles (e.g. grammatical agent, the cause of an event; grammatical patient, the affected or effected object), implemented as thematic role nodes, and organized into propositional structures through bindings with propositional nodes, sequence nodes, etc, again using binding nodes. The Concept Layer is clearly designed to model a role-filler and more classical approach to the Language of Thought.

One last example that will be of special significance later on is the system by Hadley and Hayward (1997). It contains two layers of nodes: an input layer that represents lexical items to be processed, and an output semantic layer that very closely resembles the Concept Layer of the architecture by Hadley and Cardei (1999). That is to say, the output semantic layer in the system by Hadley and Hayward (1997) uses concept nodes, thematic role nodes, and propositional nodes — bound together using binding nodes into structures — to represent the appropriate propositional semantics of a sentence being processed word by word. Hadley and Hayward showed that their architecture achieves a strong level of semantic systematicity from training on a relevantly sparse training corpus. The significant difference from the work of Hadley and Cardei is that the system by Hadley and Hayward uses an error-unsupervised Hebbian training method, and more importantly, uses *only* connectionistic processing within the atypical architecture (whereas Hadley and Cardei resorted to using a message passing algorithm in conjunction with the connectionistic processing).

2.8 Analysis of Atypical Connectionist Networks

While all four of the atypical connectionist networks surveyed above have advanced the state of the art, each has weaknesses that we should be aware of. To begin, let us take a look at the network studied by Hadley et al. (2001), and used as sub-modules by Hsiao (2002). These networks were designed to predict the semantic features of the next word based on the semantic features associated with the words processed so far (the next-word

prediction task). At first glance, it is not clear how the ability to merely predict the next word, or the meaning of the next word, figures into understanding how the meaning of a *whole sentence* is formed. Of course, this is a criticism that could apply to the typical connectionist networks examined above as well, and it is instructive to examine how they might reply to such a criticism.

At least for the SRN and SRN-like networks, e.g. by Elman (1991), Brakel and Frank (2009), and Tong et al. (2007), a response to the criticism may be that the representation in the *hidden* layer is meant to represent the meaning of *all* the words processed so far. The meaning of the next word that is inferred, and which is stored in the output-layer, is just an artifact of using the meaning of the next word — which is obviously available during training — to learn the hidden layer representations of whole sentence meanings.

This reply, however, is highly problematic when applied to the network by Hadley et al. (2001), or the sub-modules from the network by Hsiao (2002), for the reason that they use, as the hidden layer, three separate groups of neurons that merely copy each other: specifically, group one copies the activations of group two, which copies the activations from group three, which is the only group whose neural activation levels are "active" in the sense that it is the result of processing the input-layer via a set of adaptive connection weights. In such an architecture, the complete meaning of a sentence can be represented only if the sentence has three words or less. Obviously, the number of groups, N, of neurons in the hidden layer is theoretically arbitrary, but no evidence is presented that the architecture can continue to perform admirably while scaling up to a larger N. Further, the "hard" cut-off of having at most *exactly* N word meanings stored may seem cognitively questionable. In any case, it would be quite surprising, from a theoretical linguistic standpoint, that the meaning of a phrase, however complex, is simply the otherwise unstructured sequence of lexical meanings of the lexical items in that phrase.

As for the case of the network studied by Hadley and Cardei (1999), we should first note that the output Concept Layer is meant to represent meaning through displaying neural activation in certain neural units, while other units display no activation. The units with activation represent semantic concepts, semantic themes, or propositional structural relations, and are in fact semantic markers of the kind described in section 2.6. Being semantic markers, the Concept Layer can be seen as a form of Semantic Markerese into which English sentences are translated, and they therefore face the difficulties regarding Markerese raised in section 2.6 (see that section for details): namely, that doing such a translation

from English sentences to Semantic Markerese is unsatisfying in that given the semantic markers, we are not much further along in discerning the *cognitive* significance or *meaning* of the sentence presented, since Markerese requires its own syntax and semantics. Fodor has even argued that *all* morphologically simple lexical concepts are innate (Jackendoff, 2002, p. 334), an argument which may be taken as an example of how difficult it must be to make semantic markers fit within an empirical and naturalistic framework of the mind.

A possible response to the Markerese criticism is to argue that the semantic representations are actually simplifications of perceptions — "simplifications" both in the sense that it results in an incomplete but simpler model, and that the representations are generated via neural network processing of high dimensional input perceptions of the external environment (e.g. retinal signals from ocular irradiation) into lower dimensional representations. This is indeed a valid response in the cases where the output-layers use semantic features that are sufficiently fine-grained observable properties (e.g. HAS-WEIGHT, IS-RED, if not even finer-grained properties), for dimensionality reduction using neural networks is a well established process in machine learning. For example, the networks proposed by Elman (1991), Hadley et al. (2001), and Hsiao (2002) use such fine-grained semantic features in the output layer of representation (and is subject to learning), but unfortunately, the network proposed by Hadley and Cardei (1999) uses fine-grained semantic features only in an internal layer with both the mapping between concepts and features, and the connection weights used to enforce that mapping, *hard-wired* into the design[20] (ibid., p. 223) and not subject to learning in the model. Of course, one may argue that the Markerese *is* the lower dimensional symbolic representation of the high dimensional input perceptions, but no evidence is provided as to how such symbolization is supposed to occur in a connectionist manner that results in representations that have the familiar propositional structure with properly assigned thematic roles (e.g. agent, patient).

Recall also that the network proposed by Hadley and Cardei (1999) used a message passing algorithm in conjunction with the connectionistic processing to achieve strong systematicity. The reliance on a message passing algorithm seems problematic, given that no explanation is given as to how it might be possible, or if it is even at all possible, to replace

[20] "Admittedly, this is an over-simplification [in the model of Hadley and Cardei] (as indeed is [their] set of features). As frequently occurs in cognitive modeling, [Hadley and Cardei] have chosen to explore certain complexities in [their] model while simplifying others" (Hadley & Cardei, 1999, p. 223), and in this sense the model proposed in this thesis is no different. That does not, however, preclude pointing out the weaknesses of any model (including what is proposed in this thesis).

the message passing classical computation with connectionistic computation. Thus, it seems questionable as to how the message passing protocol could be instantiated neurobiologically. As mentioned previously, connectionist modelling is appealing in part for its purported neurobiological plausibility, so the reliance on the message passing non-connectionistic classical computation seems problematic when aiming towards greater neurobiological plausibility.

Given the weaknesses of the network by Hadley and Cardei (1999), the network studied by Hadley and Hayward (1997) is clearly advantageous for its ability to display strong semantic systematicity for the simplified language grammar they studied, and for its use of purely connectionistic processing. It has, however, a startling weakness that is best understood in relation to a criticism advanced by Aizawa (1997a), which is a devastating criticism if it is found completely persuasive (fortunately for Hadley and Hayward, I will argue that the criticism's conclusion proves to be a case of overstatement, although I will argue a variant form of the criticism still applies), a criticism that also applies to all the above surveyed atypical connectionist networks.

2.8.1 Explaining versus Exhibiting Systematicity

Aizawa (1997a) argues that although the connectionist architecture proposed by Hadley and Hayward (1997) is capable of exhibiting systematicity, the architecture fails to furnish readers with an *explanation* for why *human* thought is observed to be systematic. Aizawa takes a page out of the history and philosophy of science to argue that Hadley and Hayward's architecture is essentially a cognitive science version of Ptolemaic theories of planetary motion, in that both require a number of parameters to be hypothesized to be "just so" in order for the model to provide a good fit with the observed data.

The "just so" parameters (i.e. the *auxiliary hypotheses*) for the Ptolemaic theories are the epicycles and deferents used to explain retrograde motion of the planets. For the connectionist model proposed by Hadley and Hayward (1997), the auxiliary hypotheses to explain systematicity are the particular form of Hebbian learning used, the particular level of activation each output layer node is set to during training, the particular amount of activation level decay that is stipulated, etc.

One of Aizawa's points is that these auxiliary hypotheses are posited only so that the model will better fit observations, when in fact nothing *necessitates* the auxiliary hypotheses to be "just so" the way they are. "[I]t is not enough for Connectionist models to save the

phenomena; they must save the phenomena *in the right way*" (Aizawa, 1997a, p. 39), where "right way" means for Aizawa that the phenomena is the "right kind of necessary consequence" (ibid., p. 49) of the hypothesized mechanism, and not merely a result of contrivances added to a theory to make it fit the data. For example, it is just as easy to imagine a network without (or with some other arbitrarily chosen amount of) activation level decay as it is to imagine planetary motion without (or with some other arbitrarily chosen number of) epicycles, and so assuming these parameters are part of the respective model does not really provide the model with greater *explanatory power* — even if it does make the model *fit* the observed data better. Following this line of reasoning, Aizawa (1997b) goes so far as to argue that connectionism, *in general*, is unable to furnish the right kinds of explanation required to explain systematicity[21] — put differently, Aizawa argues the level of explanatory power of connectionist models (to explain systematicity in human thought) is exactly *zero*.

Hadley persuasively defends the explanatory power of the connectionist model proposed by Hadley and Hayward (1997) from Aizawa's criticism, in part, by noting the importance of having a background scientific theory as the context in which a model operates, and that given such a context, hypothesized cognitive structures can avoid the charge of being an ad hoc addition to "save" the theory (Hadley, 1997, p. 573-4). In the case of the connectionist model proposed by Hadley and Hayward, Hadley suggests "we do have such a background theory, viz., some preferred version of the theory of natural selection" (ibid., p. 574). Hadley also maintains that the model by Hadley and Hayward "posits specific structures which... might have evolved naturally... [And although] the nodes and links in [their] model must be regarded as abstractions which emerge from lower-level structure and processes... [Hadley] see no reason in principle why the requisite lower-level structures and processes could not be the product of natural selection" (ibid.). If Hadley's contention is correct, then we should accept that at least some connectionist models are in principle capable of furnishing the right kind of explanation for phenomena like the systematicity of thought through conjectured cognitive structures that might have evolved through natural selection.

[21] It should be noted that Aizawa goes so far as to argue that even Classical theories are unable to furnish the right kinds of explanation required to explain systematicity, and it would seem Aizawa (2003) continues to hold this line of reasoning.

Hadley, however, does not provide any more details as to how these conjectured cognitive structures might have *actually* evolved through natural selection — although he did not need to for the purpose of defending the model proposed by Hadley and Hayward (1997) from the critique of Aizawa (1997a), because he was simply stressing a possibility that Aizawa ignored, and further because he was stressing a possibility for a potentially *non-human* cognitive agent existing in a possibly *different* world (R. F. Hadley, personal communication, 30 June 2011). Recall from chapter 1 that, in this thesis, we are actually interested in understanding *human* cognitive processes. Thus, in lacking a proper evolutionary account of how specific cognitive structures came to be for human cognitive agents here on Earth, the model proposed by Hadley and Hayward has only some explanatory power (to explain systematicity in humans[22]), but the level of explanatory power seems to be not very satisfying since an evolutionary explanation is not provided in any amount of detail as to how any specific cognitive structures came to be for human cognitive agents (especially since evolution is specifically named as a possible background scientific theory in which the model operates in the context of)[23]. Granted, I am not aware of any attempts by *any* connectionist researchers to provide a proper evolutionary account of how specific connectionist processing modules, presumably used to model some specific cognitive structures or processes, came to be.

In the end, we see that Aizawa's contention that connectionism is completely incapable of providing an explanation to systematicity is overstated, since there is in principle a theoretical explanation (i.e. evolution) for the connectionist architecture to be the way it is. On the other hand, although Hadley's reply with natural selection is sufficient for his purpose of defending the architecture proposed by Hadley and Hayward (1997) for a possibly *non-human* cognitive agent existing in a possibly *different* world, it is not very satisfying for our purpose of studying *human* cognitive agents here on Earth, since, by not providing

[22]Or in any *specific* species of organism, actually. But note, again, that Hadley only claimed to offer an explanation of systematicity "in some *possible* cognitive agent" (R. F. Hadley, personal communication, 24 June 2011) that might not necessarily be human (ibid.).

[23]Overall, the emphasis that should be read from this argument here should be on the *gradation* of levels of explanatory power (of various theories, models, or architectures) to explain phenomena describable at varying levels of detail. An explanation of a phenomenon described at a higher level of abstraction (i.e. with less specific, particular, or contingent details) may not be very satisfying (i.e. may be found wanting in explanatory power) when the phenomenon is described at a lower level of abstraction (i.e. with more specific, particular, or contingent details). To some philosophers and computer scientists, for example, the difference in detail is what differentiates discussions in philosophy from computing science (Schulte, n.d.) — and cognitive science is an interdisciplinary field that includes parts of the computing science field.

a proper evolutionary account, it lacks the specific scientific details necessary to make the defense more than merely philosophically acceptable. A variant form of Aizawa's criticism would, therefore, still apply to the connectionist architecture of Hadley and Hayward[24]: namely, it is missing the specific contextual details that would justify the architectural decisions made as resulting in an architecture that is scientifically explainable or justifiable for the purpose of explaining systematicity in humans.

To reply to this variant form of Aizawa's criticism, we could simply provide an actual evolutionary account of how the architecture proposed by Hadley and Hayward (1997) came to be that way, over some evolutionarily significant timescale. This is not the project I want to take on in this thesis, however. Another way to provide the sufficient contextual details would be to ensure the connectionist model is adequately supported by other cognitive or neurobiological theories, such that it helps to begin a process of unifying explanations of phenomena from various fields. After all, the problematic aspect of the model proposed by Hadley and Hayward, in terms of the variant form of Aizawa's criticism, is not that it has no basis to explain its architectural decisions (recall Hadley suggested its basis is natural selection), but rather that the scientific explanatory "distance" from the architecture to its proposed basis is just too big (i.e. an evolutionary account is missing). It could be preferable, for example, to lean on results from other cognitive or neuroscientific theories, and then let *those* theories be explained by others through some evolutionary account[25]. Hadley and Hayward, however, do not claim cognitive plausibility for their model's particular structure (ibid., p. 34), and they also took an approach that is not biologically motivated (ibid., p. 35).

---

[24]The analysis of Aizawa (1997a, 1997b) can also be applied to the work of Hadley and Cardei (1999), and all the other above surveyed atypical connectionist networks as well, and the discussion here would apply equally.

[25]This idea of *sharing* the responsibility of explaining certain phenomena is not revolutionary, of course (see, e.g. Hardwig, 1991). For example, whenever the occurrence of a phenotype is explained in terms of genes, there is a tacit connection with the field of evolutionary genetics. Also, by building a model that leans on multiple theories, where each theory describes a possibly different field of phenomena, the model then represents an attempt to unify parts of those theories, which is a kind of scientific work that is worth pursuing. "Scientific understanding is, after all, a complicated affair; we should not be surprised to learn that it has many different aspects. Exposing underlying mechanisms and fitting phenomena into comprehensive pictures of the world seem to constitute two important aspects. Moreover... we should remember that these two types of understanding frequently overlap" (Schaffner et al., 1999, p. 39). It is, therefore, promising to note that evolutionary explanations of certain aspects of conceptual semantics is already being studied in the field of evolutionary linguistics (see, e.g. Jackendoff, 2002, chapter 8), and evolutionary explanations of certain aspects of neurobiology has also been studied in evolutionary genetics (see, e.g. Oldham & Geschwind, 2005, for references to some recent work).

Given all the difficulties shown above, it seems clear it would be better if we are able to more plausibly associate words and sentences to cognitive structures that can be defended as arising from perceived situations, such that the cognitive structures and processes are more psychologically or neurobiologically plausible (a task I attempt in this thesis). It must be emphasized that the connectionist architectures analyzed thus far, and especially that of Hadley and Hayward (1997), have contributed to advancing the state of the art, and I will be building on the successes they have shown.

This last point has to be reiterated. *Even* if Aizawa is right that the architecture by Hadley and Hayward (1997) resemble Ptolemaic theories, seeming Ptolemaic should *not*, on its own, be interpreted to mean being wholly unscientific or non-explanatory (as Aizawa, 1997a, 1997b, had apparently tried to insinuate). In fact, from when it was first developed in the last two centuries before the Common Era, Ptolemaic astronomy "was admirably successful" (Kuhn, 1996, p. 68) in its predictions, "is still widely used today as an engineering approximation" (ibid.), and "[u]ntil Kepler [in the 17^{th} century], Copernican theory scarcely improved upon the predictions of planetary position made by Ptolemy" (ibid., p. 156). Ptolemaic astronomy was, therefore, *not* a bad theory because it was not scientific; on the contrary, it was scientific and the best in its day, but as more evidence was gathered, the paradigm shifted, and the day moved on.

Chapter 3

Research Question

The research problem that needs to be solved is that a processing architecture needs to be created that:

1. is connectionistic in such a way so that it can be argued as possessing an orientation towards increasing neurobiological plausibility (i.e. oriented towards increasing resemblance to biological neural networks. This point will be addressed in chapter 5),

2. can be argued as modelling linguistic meaning as imagistic-like *conceptual cognitive structures* for conceptualizing situations — a conception of meaning that has been argued for by Langacker (2008), Lakoff (1987), and other cognitive linguists (this theoretical commitment will increase the orientation towards psychological plausibility of the architecture beyond mere behavioural adequacy. This point will be addressed throughout chapter 4, but especially sections 4.1, 4.2, and 4.3),

3. models the association of the aforementioned cognitive structures with input sentences (this point will be addressed throughout chapter 4 from an architectural design standpoint, while experimental evidence that the architecture accomplishes this point will be addressed in chapter 6), and

4. exhibit at least partial semantic systematicity, which also implies strong systematicity, during experimentation, in order that the architecture be behaviourally adequate in that respect (this point will be addressed in chapter 6).

If successfully solved, the architecture promises to thus:

1. demonstrate one possible way of providing a processing model of conceptual semantics, largely as proposed in Cognitive Grammar, that can be implemented within a connectionist framework (recalling from section 2.6 that Cognitive Grammar focuses on the phenomenological aspect, while this thesis is committed to understanding the processing aspect),

2. experimentally demonstrate one possible way of exhibiting the aforementioned systematicity properties within a connectionist framework, and thus provide corroborative evidence for the proposed architecture and conceptual semantics, largely as proposed in Cognitive Grammar, and

3. answer the challenge, in terms of the variant form of the criticism proposed by Aizawa (1997a, 1997b) as discussed in section 2.8.1, for a bona fide scientific explanation of the aforementioned systematicity properties in humans.

The rationale for pursuing the proposed research problem will now be discussed below. The discussion will also further justify that the problem is previously unanswered, and that it is worthwhile to be answered in this thesis.

Keeping in mind the context behind the overall research, as explained in chapter 2, recall the overall goal of this thesis. It is to explore and model, using machine models within a naturalistic scientific framework, the human cognitive processes dealing with language comprehension. We are not in particular interested in a study of phonology, and so input sentences to the architecture may be modelled in a simplistic manner.

In modelling the human cognitive processes for comprehension of meaning, however, we must focus on understanding the nature of the cognitive structures employed within cognition, where in our case of dealing with language, the conceptual cognitive structures would be of the meaning of sentences processed. Recall from section 2.5, however, that all the typical connectionist network architectures surveyed in section 2.3 made no claims toward the cognitive structures and computational procedures, and their relationship with more psychologically realistic theories of cognition, whilst seeking greater behavioural adequacy. *The proposed architecture in this thesis, in contrast to previous work, commits to an analysis of cognitive structures and computational procedures, and relates them to conceptual semantics, largely as proposed in Cognitive Grammar.*

Given the nature of human cognitive abilities, and in particular systematicity and productivity, we may expect the conceptual cognitive structures themselves to possess some

systematicity and productivity. Whereas these conceptual cognitive structures are to be the *meaning* of sentences (in accordance to conceptual semantic theory), the proposed architecture in this thesis will associate sentences to conceptual cognitive structures (i.e. meaning) that use sufficiently fine-grained semantic features that can plausibly be defended as arising from perception (e.g. as arising neurobiologically from ocular irradiation). The proposed architecture thus stands in stark contrast to previous work: recall from section 2.8 that all the atypical connectionist network architectures surveyed in section 2.7 either associates sentences to symbolic semantic markers, or fails to capture the entire meaning of complete sentences.

Since cognition as envisioned by "cognitive linguists is noninsular, being grounded in perception and bodily experience" (Langacker, 2008, p. 28), and meaning is embodied in that it "is understood via real [human] experiences in a very real world with very real bodies" (Lakoff, 1987, p. 206), it would be better for cognitive modelling to devise a system that is connectionist in nature in order that the system be more neurobiologically plausible. This is the approach taken in this thesis, and furthermore, some specific properties of connectionism will be taken advantage of to enable an argument for greater neurobiological plausibility in chapter 5.

Given the problem of the lack of detailed analysis of the cognitive structures, if any, emerging within typical connectionist networks studied in past research (recall section 2.5), and since none of the surveyed connectionist networks implement a processing model of conceptual semantics, largely as proposed in Cognitive Grammar, it will be clear that, in contrast to previous work, an architecture that is designed to take advantage of the conceptual semantics and Cognitive Grammar research on conceptual cognitive structures increases the architecture's orientation towards psychological plausibility.

The research problem stated in this chapter is thus clearly unanswered by previous research as evidenced by the above discussion, and by the analysis of typical connectionist networks in section 2.5, and of atypical connectionist networks in section 2.8. While the atypical or hybrid connectionist architectures, and in particular the work of Hadley and Cardei (1999), of Hadley et al. (2001), and especially of Hadley and Hayward (1997), have made many advances in the area of modelling semantic representations and employing neurobiologically more plausible learning regimes, more can be done (and *will* be done in this thesis) to lessen the reliance on semantic markers, to better capture the meaning of whole sentences, and to provide greater explanatory coherence in conjunction with other

cognitive sciences in order to better explain the systematicity of thought while taking into account more empirical evidence of the functioning of cognition.

Furthermore, the proposal to encode meaning as cognitive structures from embodied experiential conceptual semantics, which has been argued for by cognitive linguists as being more psychologically plausible, is certainly not present in any of the architectures surveyed, again reinforcing the fact that the research problem stated in this chapter is thus clearly unanswered by previous research. Note that although embodiment as a property of conceptual semantics imply that any complete model of semantics must also completely model the relevant (human or robotic) bodily processes involved, for the purpose of this thesis, I am only interested in an *incomplete* model focused entirely on the processes involved in *associating* sentences and conceptual cognitive structures — this would then represent a first step in the direction of a more complete model that is based on an embodied experiential conceptual semantics that has been argued as more psychologically plausible by cognitive linguists.

Therefore, the proposed architecture, if it is experimentally shown to successfully exhibit some degree of systematicity, will thus demonstrate one possible way of providing a processing model of conceptual semantics, largely as proposed in Cognitive Grammar, that can be implemented within a connectionist framework. If the proposed architecture successfully show experimental evidence of the aforementioned systematicity properties, it will therefore further provide corroborative evidence for the proposed architecture and conceptual semantics, largely as proposed in Cognitive Grammar. Furthermore, it would thus also be an answer to the challenge, in terms of the variant form of the criticism proposed by Aizawa (1997a, 1997b) as discussed in section 2.8.1, for a bona fide scientific explanation of the aforementioned systematicity properties in humans.

It may seem strange that the goal of this thesis is, in part, to demonstrate one possible way of providing a processing model of conceptual semantics, largely as proposed in Cognitive Grammar, that can be implemented within a connectionist framework, with a large emphasis on arguing that the proposed architecture is oriented towards increasing neurobiological plausibility so that it is at least a stepping stone on the path to a fully neurobiologically plausible implementation. Stranger still may be that the systematicity properties will be shown through experimentation on artificial languages produced by simple recursive grammars. It is, however, important to stress that quite a few connectionist

implementations are of this kind within cognitive science (an interdisciplinary field that includes the study of artificial intelligence from the computing science field), as demonstrated by all the works surveyed in sections 2.3 and 2.7, and also the works of, e.g. Stewart and Eliasmith (2009); Smolensky and Legendre (2006); Neumann (2002); Elman (1991); and Elman (1990).

Chapter 4

Approach and Design

4.1 Phonological and Conceptual Cognitive Structures

In this thesis, we will adopt a learning framework that presupposes (roughly, but we will make this presupposition much more precise in this section) that the language learner is able to discern that a sentence uttered by a teacher is *intended* to refer to the observed situation (though it need not actually refer, given our stance towards intentionality. Recall section 2.1 and section 2.6. Thus, the learner is presupposed to associate the perceptions of an utterance to the perceptions of a situation). This presupposition, an auxiliary hypothesis to the model being built in this thesis, bundles together three different issues.

Firstly, the language learner is assumed to be able to discern a particular utterance (which we ordinarily would interpret as a sentence) from the stream of auditory signals generated continuously by the ears. This assumption may be justified in our incomplete model as we are not in particular interested in a study of phonology based on auditory signals from the ears. Furthermore, "even young children, who have not yet reached the stage of producing multi-word utterances, are frequently able to obey simple imperative sentences which contain words in syntactic positions where the child has never encountered the word before" (Hadley, 1994, p. 8), thus there seems to be some evidence that even young language learners are capable of discerning sentences (as in an utterance) from the continuous stream of auditory signals. In any case, we are in good company in assuming "sentences and not words as the wholes whose use is learned" (Quine, 1960, p. 13) by young language learners: for example, we already saw the emphasis on learning T-*sentences* from

our previous discussion on truth-theoretic theory of meaning (section 2.6); and it is also commonly assumed by many connectionist researchers (including, e.g. Hadley & Cardei, 1999, and Hadley & Hayward, 1997).

Secondly, the language learner must assume that the teacher intends the sentence utterance to be associated with the observed situation. The assumption that the to-be-learned sentence "P" is *intended* to be associated with the presently observed situation is made on the presumption of a version of the *Principle of Charity*, which is a social convention that facilitates linguistic communication (variations of which have been argued for to aid communication of other forms or sources of meaning, for instance, Gricean implicatures. See Grice, 1975). We are, therefore, of course assuming that human language learners are social beings who innately value, or quickly learn to value, a Principle of Charity.

Thirdly, the language learner assumes that it is the currently observed situation, being a fragment of the entirety of the state of affairs of the world, that the utterance is intended to be associated with. We are thus investigating a kind of meaning that is closely related to what Quine called *affirmative stimulus meaning*. Consider the word "Gavagai", which is uttered by a native speaker upon seeing a rabbit, we may wish to say that the rabbit prompted the native's uttering "Gavagai" when in fact it is the ocular irradiation *stimulations*, and not rabbits, that prompted the utterance. Or consider Quine's example, where a learner asks "Gavagai?" (a one-word sentence) when a rabbit runs by the native and learner, and the native assent to the utterance: again, it is stimulations, and not rabbits, that prompts the native's assent (Quine, 1960, p. 31), for "[s]timulation can remain the same though the rabbit be supplanted by a counterfeit. Conversely, stimulation can vary in its power to prompt assent to 'Gavagai' because of variations in angle, lighting, and color contrast, though the rabbit remain the same" (ibid.). Thus, when the language learner concludes (by whatever means or evidence, including by appeal to the Principle of Charity) that the native has assented or affirmed a sentence as being intended to be associated with an intersubjectively shared situation, it is the stimulation caused by the situation in the native that is the *affirmative stimulus meaning*[1] of the sentence for the native; and it is the stimulation caused by the situation in the learner that the learner can assume is the

[1] "More explicitly... a stimulation σ belongs to the affirmative stimulus meaning of a sentence S for a given speaker if and only if there is a stimulation σ' such that if the speaker were given σ', then were asked S, then were given σ, and then were asked S again, he would dissent the first time and assent the second" (Quine, 1960, p. 32).

affirmative stimulus meaning for that sentence.

The language learner, by way of collecting from many natives under many situations a large data set of pairings between affirmative stimulus meanings and sentences (call a single such pairing an *ASM pair*), can construct a model of what affirmative stimulus meaning a sentence in the data set has for an appropriately large segment of the native's community[2]. After building up such a sufficiently large data set of ASM pairs, the learner would be able to work out the semantically significant components (e.g. words) of the sentences in the language whilst associating them to fragments of stimulations from the affirmative stimulus meanings in the data set, such that higher order relationships between fragments of stimulations can also be worked out so that how affirmative stimulus meanings compose or combine in the data set is also worked out. From here, the language learner can then infer the affirmative stimulus meaning for a novel sentence (so long as all the semantically significant parts of the novel sentence have been seen in the data set).

The meaning of individual semantically significant components of a sentence (e.g. words) thus depend on the ASM pairings in which they occur, just as Quine proposed that "words are learned only by abstraction from their roles in learned sentences" (ibid., p. 51). Therefore, meaning is holistic in the aforementioned Davidsonian sense: the meaning of a sentence depends on the semantically significant parts of the sentence, and the meaning of the parts depends on their systematic contribution to the meaning of *all* sentences in the language in which, to the speaker or listener, they are known to occur. The meaning of words is thus revisable, since the data set of ASM pairs is built up over time as the learner experiences more of the language being used by the native speakers. Since we know that "[i]ndeterminacy [of translation] means not that there is no acceptable translation, but that there are many" (Quine, 1987, p. 9), it is not a problem that meaning is revisable in the process of translating the native's language into the ASM pairs in one's head. Whereas a "good manual of translation fits all checkpoints of verbal behavior, and what does not surface at any checkpoint can do no harm" (ibid.), and whereas there are more checkpoints to fit the more experience one has with the native speakers' language usage, it is still the case that any model of meaning built by the language learner will always be underdetermined by all possible observations.

[2]Of course, if the segment of the community is sufficiently small, the sentence (which may be a one-word sentence) may be called "jargon" by the wider community.

There are other forms of meaning that this thesis does *not* address, given our focus on a form of meaning that is closely related to affirmative stimulus meaning. For example, *negative stimulus meaning*, meaning of *non-observational sentences*[3], Gricean implicatures, etc. The narrow focus in this thesis, however, should not be too worrisome as we know meaning is not a simple phenomenon, and furthermore, given our allegiance to naturalistic science, it seems that whatever meaning is, it must be something that resides in neurobiological processes in the brain, operating on stimulations (caused by, e.g. situations or utterances) the brain receives neurobiologically from, e.g. the eyes, ears, and other parts of the nervous systems, and on past brain states.

Talk of pairings between sentences or utterances and stimulations, may be sufficiently detailed for a theory developed for the philosophy of language, but we will need more detailed analysis if we are to develop a theory or model for cognitive scientific research. In particular, since this thesis is interested in what is occurring within cognition between cognitive structures of various types, let us be clear as to what "cognitive structure" means: *cognitive structure* is a spatial-temporal arrangement of atomic cognitive elements, which we will assume are neural activations in the brain. Two types of cognitive structures are of particular interest to us: *phonological cognitive structures* and *conceptual cognitive structures*, both of which are cognitive structures instantiated in the brain in response to different kinds of external stimulus, and will be explained, in turn, in more detail below.

To begin, consider what is meant by *sentence utterance*: a *sentence utterance* is a sequence of mechanical waves in the air that hits the eardrums of a person. Many sentence utterances will ordinarily be said to be the same when heard by a person, of course, since a wide range of mechanical wave sequences would elicit the same response from the hearer if the hearer, for example, was asked to assent or dissent to them (e.g. as, *ceteris paribus*, when the same sentence is spoken aloud in two different voices).

Let A be the specific area of the brain used for processing auditory signals from the eardrums, perhaps consisting, in part, of the human auditory cortex where there is evidence of *tonotopic representation* of auditory signals (Weisz, Wienbruch, Hoffmeister, & Elbert, 2004), then a cognitive structure P consisting of the neural activations in A (or possibly a sequence, over time, of such neural activations in A) is instantiated, possibly over time,

[3]That is, sentences whose "stimulus meanings vary over society in as random a fashion as that of 'Bachelor', and it is only the few verbal links that give the terms the fixity needed in communication" (Quine, 1960, p. 56).

each time a sentence utterance U is heard[4]. We call P a *phonological cognitive structure*, and we say that P is *prompted* by U.

Now let us reconsider what is meant by a situation: a *situation* is a state of a part of the external physical world. As previously mentioned in footnote 9 (section 2.6), unless a precise distinction is required between a situation as a whole versus a fragment of it, the term "situation" will be used for both cases without specifying the distinction. When a person observes a situation (and let us focus only on the visual sense of that person for the purpose of this thesis), a set of retinal signals will be generated due to ocular irradiation. Let V be the specific area of the brain used for processing visual signals from the retinas, perhaps consisting, in part, of the human visual cortex (see, e.g. Solomon & Lennie, 2007), then a cognitive structure C consisting of the neural activations in V (or possibly a sequence, over time, of such neural activations in V) is instantiated, possibly over time, each time a situation S is seen[5]. We call C a *conceptual cognitive structure*, and we say that C is *prompted* by S.

If the situation S consists only of a spatial arrangement of physical elements, then the resultant conceptual cognitive structure generated via ocular irradiation due to that situation will be said to be *static* (the process that transforms ocular irradiation to static conceptual cognitive structure has been studied in neuroscience, and also in connectionist research, particularly by those employing Deep Belief Nets. See Bengio, 2007, p. 2-3). If the situation involves a temporal element, that is to say, if the spatial arrangement of physical elements evolves from time t_0 to time t_n, then the resultant cognitive structure generated in V from time t_0 to t_n is said to be a *dynamic* conceptual cognitive structure.

At the beginning of this section, we said we presume *roughly* that the language learner is able to discern that a sentence uttered by a teacher is intended to be associated with the observed situation. We refined this presupposition in terms of pairings between sentences and affirmative stimulus meanings. Now that we have developed even more refined terminology, we can say with more precision just what the language learner is doing: the learner

[4]Note that this is an extremely simplified abstraction that will have to be refined through further empirical neuroscientific research, perhaps specifically of the human auditory cortex. Having said that, any comparison of the brain or this abstraction with modern computers might be misguided, as such a comparison would have to first justify how the brain can be compared with a *von Neumann architecture*.

[5]Note that this too is an extremely simplified abstraction that will have to be refined through further empirical neuroscientific research, perhaps specifically of the human visual cortex. Again, any comparison of the brain or this abstraction with modern computers might be misguided, as such a comparison would have to first justify how the brain can be compared with a *von Neumann architecture*.

is associating phonological cognitive structures with conceptual cognitive structures. Notice this is *exactly* the key fundamental insight from conceptual semantics as posited within Cognitive Grammar theory (recall section 2.6), and thus "conceptual cognitive structure" is synonymous with "conceptualizations" as posited in Cognitive Grammar. Phonological and conceptual cognitive structures are generated neurobiologically, respectively from perceptions of sentence utterances and perceptions of situations, and thus all of our above discussions (e.g. on the Principle of Charity, on how meaning of words are learned from ASM pairs) in this section continue to apply.

Since we are now talking about pairings between cognitive structures, and thus theorizing at the cognitive level rather than the social level or level of intersubjective agreement of stimulations, the association between the cognitive structures can be described in greater detail. Specifically, the association between phonological and conceptual cognitive structures is performed through psychological processes that have been identified (Langacker, 2008, pps. 16-7) as basic and evident in many facets of cognition, namely, *association*, *automatization*, *schematization*, and *categorization*.

With that in mind, a connectionist architecture may provide a plausible model of the association between the phonological and conceptual cognitive structures since connectionistic learning is clearly *associative*; supports *automatization*, since the repeated supervised learning of a teaching pattern and target pair can *entrench* that pair so that the observation of one would automatically generate the other; supports *schematization* in that artificial neural networks are known to be fairly good at extracting commonalities amongst patterns in a data set of multiple experiences; and supports *categorization*, where novel experiences are interpreted in terms of the categories learned by the latent variables in the hidden layer's neural units in a neural network, through the past experiences in its teaching data set (that connectionism supports the psychological processes used in Cognitive Grammar is in fact acknowledged by Langacker. e.g. see Langacker, 2008, p. 10, and Langacker, 1991, p. 533).

Therefore, the meaning of a sentence utterance[6] for a person is the meaning of the phonological cognitive structure generated in that person from hearing that sentence utterance, and is its properly psychologically associated conceptualization (i.e. the conceptual

[6]As already stated, this thesis will only deal with observational sentences that have at least an affirmative stimulus meaning. Other kinds of meanings that are possibly based on a recombination or extension of affirmative stimulus meaning will not be considered in this thesis.

cognitive structure that would be generated by an instance of the affirmative stimulus meaning of that sentence utterance for that person). In a sense, we are thus translating from a sentence utterance to a phonological cognitive structure, and finally to a conceptualization. Notice, however, that sentences are not being translated into an artificial language that itself requires its own interpretive semantics for us to make sense of it. Rather, sentences are being mapped, in a sense, to the associated conceptualization, which must be *embodied* for a complete cognitive model[7].

Embodiment means that the "[c]ognitive models... are not made up merely of items in an artificial language. In experientialist semantics, meaning is understood via real experiences in a very real world with very real bodies" (Lakoff, 1987, p. 206). Thus, conceptual cognitive structures — being generated neurobiologically from stimulations (e.g. ocular irradiation) from a situation, which a sentence would *traditionally* be said to be *intended* to (though need not successfully) refer to — do not themselves *have* a formal interpretive semantics that interprets symbols to things (e.g. "rabbit" to a rabbit), the way a Fodorian Language of Thought does, but rather it *is* the semantics for that sentence, such that the semantics is itself in non-semantic, non-intentional, and naturalized terms that can be studied as the causal neurobiological effect of the physical situation.

This last point may be contentious, and a complete argument for it would take us far afield, but it does demand a detailed sketch to show that it is a viable option *methodologically*, at the very least, as it impinges on our interpretation of what the connectionist architecture built in this thesis is intended to model (this sketch will be based on the same argument advanced by Jackendoff, 2002, pps. 278-9). To Fodor and Pylyshyn, "Classical theories — but not Connectionist theories — postulate a 'language of thought'... they take mental representations to have a combinatorial syntax and semantics" (Fodor & Pylyshyn, 1988, p. 8). The mental representations are the Semantic Markerese that has been discussed in section 2.6. Fodor and Pylyshyn notes that representational states are intentional (ibid., p. 4), meaning they are *about* something, namely entities in the external physical world. Suppose an English sentence has been translated into Markerese, which is itself a classically symbolic structure that can be interpreted to *represent* something in the world. The fact that, after translating the English into Markerese, we are still left with a structure that

[7]But recall that our goal in this thesis is for an *incomplete* model or simulation, as we will not be creating a model complete with a human or robotic body. Also see chapter 3.

itself requires an interpretation, i.e. a *formal* semantics, is intensely troubling.

One reason that it is troubling is that "we can know the Markerese translation of an English sentence without knowing the first thing about the meaning of the English sentence: namely, the conditions under which it would be true... Translation into Latin might serve as well" (Lewis, 1970, p. 18), since after translating into Latin, we would still be left with the same problem of having to come up with a second, "auxiliary", semantics (for Latin, in this case). The mention of Latin here is not meant to be facetious, as Fodor believes that the Language of Thought, which he sometimes call *Mentalese*, in many important ways resemble natural languages, saying that, e.g. "for most present purposes, one might as well assume that English *is* Mentalese" (Fodor, 1999, p. 513, emphasis in original), but that "English is [either] not the language of thought, or that, if it is, the relation between syntax and semantics is a good deal subtler for the language of thought than it is for the standard logical languages" (Fodor & Pylyshyn, 1988, p. 18). What is clear is that Mentalese *itself requires* a semantics, a theory of which must explain how intentional expressions in the mind can make contact with the physical things they are about.

Information-based semantic theories (IBST) — which analyzes meaning in terms of information, which can be used to causally develop associative, neurobiological associations between mental structures and real world experience — purport to provide just such a theory, but Fodor (1999) has provided a thorough analysis of how IBSTs have a "terrible problem" (ibid., p. 515)[8]. In the same article, Fodor tries to rehabilitate IBST by arguing for his solution to the "terrible problem", only to end up being "far from sure that the [solution] is right" (ibid., p. 513), which seems to show just how providing a theory of formal Mentalese semantics is not easy at all.

To be fair, the same question can be asked of conceptual cognitive structures that are associated with phonological cognitive structures. Namely, how does a conceptualization (i.e. conceptual cognitive structure) in the head make contact with the physical things they are "about" (note the scare quotes) in a way that can be explained using only non-semantic, non-intentional, and naturalized terms? If this can be answered successfully, then we may say that natural language is understood in terms of phonological and conceptual cognitive structures, but the conceptual cognitive structures do not themselves have an additional,

[8]The "terrible problem" is technical and specific to IBSTs, and so details of it are not really relevant here. In brief, it is that IBSTs are "demonstrably wrong... about representational uses of symbols" (Fodor, 1999, p. 516).

secondary, auxiliary semantics. To briefly sketch out how this can be done, we recall the possibly radical step taken at the end of section 2.1: at least within our scientific methodology, we will take conceptual cognitive structure "just as pure non-intentional structure, as we did (less controversially, [Jackendoff] hopes) with phonology and syntax" (Jackendoff, 2002, p. 279). Rather than incorporating a theory of intentionality within the model being developed in this thesis, given that even Fodor (1999, p. 513) acknowledges that intentional and semantic symbols are *not* compatible with naturalistic scientific theories[9], the problem of reconstructing a notion of intentionality will be left as a wholly different project, one that has been tackled by, e.g. Jackendoff (2002, Chapter 10) and Dennett (1987a).

Having committed to a non-intentional conceptual cognitive structure, we now need to sketch out how it may be produced within a connectionist architecture. As previously mentioned, the conceptual cognitive structure is capable of being causally created in the learner's brain through the neurobiological processing of perceptions and stimulations of a situation — a process that has been studied in connectionist research, particularly by those employing Deep Belief Nets (e.g. Bengio, 2007, p. 2-3). In the case of Deep Belief Nets (DBN), for example, they can in fact learn a higher level neuro-encoding of perceptual features in an entirely unsupervised manner, using error-unsupervised learning algorithms. This means DBNs can learn without using error-backpropagation, using Hebbian-style training, which is "widely believed... [to be] closer to biological reality than the commonly used method of backpropagation of error" (Hadley & Cardei, 1999, p. 7).

More importantly though, DBNs can learn in an unsupervised manner using unlabelled data — i.e. it can learn without a predetermined "target" being associated with each input — and has been shown to be capable of learning to represent visual input with a deep hierarchy of nonlinear feature detectors (Hinton, 2010). Just to serve as an example, contrast this with the network by Hadley and Cardei (1999), where "the entire target meaning representation of the sentence being processed is active throughout the processing of a given input sentence" (ibid., p. 22) during network training, meaning the input is labelled with a target that is provided from the data set the model learns from, such that the structure used for representing meaning of sentences assumes meaning has a propositional character (note that the approach of using labelled training data will be used in this thesis

[9]This is a position that others have argued for as well, including Jackendoff (2002), Dennett (1987a), and Quine (1960). Of course, other philosophers have disputed this position, including Lynne Rudder Baker (R. F. Hadley, personal communication, 24 June 2011).

as well[10]). R. F. Hadley (personal communication, 19 May 2011), instead of describing the data set as being "labelled", describes the data set as a set of input sentences where each input sentence as a whole is associated with a "guessed" meaning, which need not always be accurate. This can be put differently and instead be described as each sentence, as a whole, being labelled with a target (where a target is a vector, for example, if meanings are interpreted as vectors in a vector space), and the architecture builds a regression model with the labelled data through supervised learning (regression from a data set that may contain noise or mislabelled data points does not change the fact that it is still supervised learning).

Similarly, the connectionist architecture (hereinafter called *S11*) to be developed in this thesis will also learn with labelled data in a supervised manner similar to the work of Hadley and Cardei (1999), and Hadley and Hayward (1997). The point in discussing the capabilities of DBNs is to show that there is already some existing research that is suggestive of how something like the conceptual cognitive structure might be formed experientially, without intentionality, and in a possibly neurobiologically realistic manner (e.g. there is evidence that certain DBN architectures work in a manner similar to the standard model of the visual cortex. Hinton, 2010, p. 183), although some researchers find that contention to be highly contentious and believe other ingredients are necessary for "meanings" to arise.

To be sure, I am writing a very large cheque that will have to be cashed on some future date, and the arguments above do not constitute a knock-out blow against a theory of cognition that includes a Markerese that itself has an intentional interpretive semantics. The difference from writing a cheque for Markerese, however, is that the account the cheque will be drawn against is currently already in the process of being filled by naturalized scientific connectionist research that has produced interesting computational models that can perform interesting tasks (e.g. image classification, image generation). This shows, methodologically at least, that what is urgently necessary to inform the debate is more investigation on connectionist models of conceptual cognitive structure.

[10]It should be noted that the architecture developed in this thesis is specifically designed to hopefully provide a way forward in terms of integration with other modules that may have a DBN-like architecture, since the structure used for modelling meaning of sentences assumes meaning has an imagistic-like character that uses sufficiently fine-grained semantic features that can plausibly be defended as arising from perception. This point will not be further pursued as it is an issue for future research.

4.2 Scene Structure — Modelling Conceptualization

Recall that Cognitive Grammar's study of conceptual semantics envision meaning as conceptualization, which can be approached from either a phenomenological or a processing standpoint: i.e. we can attempt to characterize either our mental experience per se or the processing activity that constitutes it. Conceptualization, characterized phenomenologically, broadly includes various kinds of mental experiences, such as sensory, kinaesthetic, and affective experiences that "develop and unfold through processing time [the real time used to fully apprehend the experience] (rather than being simultaneously manifested). So even if 'concepts' are taken as being static, conceptualization is not" (Langacker, 2008, p. 30). So rather than being of a propositional character, semantics is conceived of as being imagistic (although "image" is to be read in a broad way to include all kinds of senses, rather than just the visual sense), but where the imagistic experiences are schematized into structures containing features common to wide classes of experiences (ibid., p. 32). Langacker suggests that these features may contain notions of line, angle, curvature, brightness, focal colours, contrast, boundary, etc. (ibid., p. 33).

In this thesis, we are interested in the processing standpoint of characterization of conceptualization, and specifically we model it with a connectionist architecture that is hoped to be oriented towards increasing neurobiological and psychological plausibility. Recall that conceptualizations are cognitive structures, and a cognitive structure is a spatial-temporal arrangement of atomic cognitive elements, which we will assume are neural activations in the brain. We can thus model conceptualizations with a connectionist architecture, insofar as connectionist architectures can plausibly model the processing of neural activations. The part of the connectionist architecture S11 being developed in this thesis that models a conceptualization is called the *scene structure*.

Since we would be satisfied with an incomplete model (as discussed in chapter 3), and since many of the schematized features, from imagistic experiences, that would be a part of a conceptualization are extracted through processes that are the central problems in the fields of biological and computational vision, that level of processing detail will be simplified and abstracted out of the present scene structure model. In fact, although conceptualizations include mental experiences that are dynamic and experienced through the real time used to fully apprehend the experience (i.e. as opposed to the entirety of the experience being manifested simultaneously), as a simplifying assumption for this thesis, the scene structure

will only model static conceptual cognitive structures (i.e. structures that are manifested simultaneously).

Described phenomenologically, the scene structure models a simplified form of a *profile* in Cognitive Grammar, where a *profile* is a particular substructure of a body of conceptual content that has been selected as being the "general locus of viewing attention" (Langacker, 2008, p. 66). Specifically, the scene structure model will capture only the following: two kinds of *conceptual entities*, namely that of *things*, and *relationships*; in regards to relationships, a simplified form of *process* relationships, and *non-processual* relationships will be included; and lastly, again regarding relationships, the additional classification of *trajector*, *path*, and *landmark* are also modelled.

To be clear, a *thing* is a conceptual cognitive structure that is the product of grouping or reification, resulting in a single entity suitable for higher levels of conceptualization. For instance the grouping of ocular irradiation stimulations, through neurobiological processing, of several independent bright dots in the sky as a single conceptualization of the Big Dipper (but recall that it is stimulations, and not the Big Dipper, that prompts the conceptualization). As another example, the reification of ocular irradiation stimulations, through neurobiological processing, of seven black dots on the printed page as an instance of a hexagon[11]. In this broad and phenomenological sense, Cognitive Grammar proposes that, generally, *a noun profiles a thing* (Langacker, 2008, p. 106).

A *relationship* between things, or other conceptual entities, can be entertained conceptually as the placement or configuration of all those entities. Those entities may either be fully manifested in cognition simultaneously (i.e. atemporally), or they may manifest and transform over *processing time*. *Processing time* is the real time used to fully apprehend the entities' changing set of relationships (i.e. state) over time (i.e. temporally). The atemporal relationships between entities fully manifested in cognition simultaneously are referred to as *non-processual* relationships. The temporal relationships between entities that

[11]The hexagon example is due to Hadley (2009, p. 1207). In citing Hadley here, I fully acknowledge the theoretical possibility that cognition may require some form of consciousness (or perhaps consciousness is an emergent property that is completely supervenient on the underlying physical processes). It is, however, *not* claimed that *linguistic* understanding in the form presented in this thesis, being similar to affirmative stimulus meaning, require consciousness. Certainly, consciousness (and most certainly a consciousness that requires the classical conception of intentionality) is not being considered in the model being built in this thesis, in order to simplify and reduce the amount of ontological "stuff" this thesis must commit to existing.

manifest and transform over processing time are referred to as *processes* or *processual* relationships. Cognitive Grammar argues that, generally, *a verb profiles a process* (Langacker, 2008, p. 112), whereas traditional categories of adjective, adverb, and preposition are subsumed into the Cognitive Grammarian's category of non-processual relationship, although they can be further subcategorized along other lines (Langacker, 2008, p. 100).

One useful line along which relationships can be categorized is in how the participating conceptual entities are chained together. When a relationship is construed, its participating entities are usually seen at differing levels of prominence or salience. The thing that is of primary focus is called the *trajector*, while the thing with secondary focus, if any, is called the *landmark* (Langacker, 2008, p. 113). Furthermore, the relationship is often distinguished also by the *path* along which the trajector "moves", whether physically or metaphorically, through to the landmark, when characterized phenomenologically. For example, as Lakoff has shown (in extending the analysis of Lindner and Brugman), the trajector/path/landmark features of a relationship are a part of the cognitive construal of nearly one hundred kinds of uses of the English word "over" (Lakoff, 1987, p. 418).

The trajector, path, and landmark features allow us to categorize some non-processual relationships as being that of the traditional adjective, adverb, or preposition. Namely, those non-processual relationships that have a thing as its trajector are profiled by what is traditionally called an adjective. Similarly, those with a trajector that is a relationship (processual *or* non-processual) are said to be profiled by an adverb. Lastly, prepositions profile those with both a trajector and a landmark, but whereas the trajector can be any conceptual entity, the landmark must be a thing. (Langacker provides good arguments for these categorizations: see Langacker, 2008, pps. 113-7) There is overlap between these categorizations, whereas the traditional grammatical classes were viewed as mutually exclusive, but "this overlap is one reason for thinking that the traditional categorization. . . is less than optimal" (Langacker, 2008, p. 117).

Finally, entities may also have various other features that do not necessarily arise through neurobiological processing of perceptual stimulations, features that Jackendoff has categorized as *valuations* (Jackendoff, 2002, p. 312-3). The trajector/path/landmark distinction would, in this scheme, be one type of valuation. In particular, "these features register not the perceptual qualities of an entity but the associated 'feel,' so to speak. I called these the 'affects' or 'valuation' of the percept" (Jackendoff, 2002, p. 312). Other features that are

valuations include: *self-produced* vs. *non-self-produced*, used to distinguish between hallucinations from voluntary imagery; and *internal* vs. *external*, to distinguish between the result of experiencing perceptual stimulations of something in the world, as opposed to images not resulting from direct perceptions (Jackendoff, 2002, p. 313). Langacker suggests some other distinctions that would fall under this category as well, including *focal prominence*, the degree of salience in an entity, as it is a "conceptual phenomenon, inhering in our apprehension of the world, not in the world per se" (Langacker, 2008, p. 72-3); and also includes the level of *subjectivity* vs. *objectivity*[12] of an entity, that is the relative phenomenological "closeness" or distance (but not necessarily in a spatial sense) of an entity to either the conceptualizers or the conceptualized (Langacker, 2008, p. 260).

Although the described conceptual categories above do not encompass all the traditional grammatical categories, they do form a working subset of Langacker's Cognitive Grammar theory, and will be the subset used for the purpose of phenomenologically characterizing the cognitive modelling implemented by the S11 architecture built in this thesis. Further, although these categories are associated here with their traditional grammatical categories, *within the connectionist architecture S11, these associations will not be "hand-built" into S11, but the S11 network must learn for itself how to make the proper associations from the training data set.* The identification of these conceptual categories with their traditional grammatical categories is presented here for expositional purposes only, so that the reader can more easily grasp what the connectionist system will be tasked to learn, and what it models from the phenomenological standpoint of cognition.

[12]The use of "subjectivity" and "objectivity" here is in reference to the phenomenological nature of the entities within a *conceptualization*, and is used to distinguish between that which is construed as subjects (e.g. speaker, audience member) having the (subjective) *experience* of conceptualizing, versus that which is being construed (objectively) as a separate entity from the subjects. These terms are not being used to refer to the grammatical notions in categorizing *words* or *phrases* of words.

4.3 Scene Structure as a Connectionist Model

Within the connectionist system S11, we will be modelling each conceptual entity as a sequence[13] of neural activations (call such a sequence an *entity-rep*[14]). The encoding scheme will essentially be feature based, so that if an entity has three features, then the sequence of neural activations encoding that entity will have activations in the three neural units encoding those three features (we may interpret varying levels of activations in a unit as a strength of recognition of the presence of that feature). More details on the encoding scheme, especially in regards to the features to be modelled in the things, relationships, and valuations, will be presented in section 4.4.

Please be reminded that, except for valuations, features of things and relationships are, in a complete model (though we are currently building an incomplete model as a first step), the conceptual features extracted from the perception (e.g. through ocular irradiation) of a whole or part of a physical entity, possibly through time; these conceptual features may be "low-level" (e.g. retinal signals from the rods and cones of the left eye) or may be "higher-level" schematizations (e.g. a horizontal line in a particular part of the visual receptive field). In the case of valuations, the features may be extracted from other embodied processes, e.g. through our kinaesthetic senses, that do not directly "map" out to real physical entities independent of the speaker's *body* (this is not to say that the features map out to *nothing* at all, but rather that they may simply map to relationships or processes previously abstracted from relationships or processes internal to the speaker's body that were prompted during past experiences)[15]. These extractions may, perhaps, proceed through some employment of Deep Belief Nets (Bengio, 2007, p. 2-3), for instance. Here we are only interested in the structure of the extracted features, as analyzed through Cognitive Grammar, and its connection to sentences through connectionistic means.

Since non-processual relationships involving a thing or a relationship in essence modify or reshape that thing or relationship in terms of the features constituting it, we may view

[13]We will reserve the word *vector* for mathematical vectors, and instead use the sequence abstraction as the abstract data structure for encoding neural activations from what would traditionally be termed a *layer* of neural units in a feedforward network.

[14]Recalling our discussion from section 4.1 and other sections on intentionality and representation, please note that the term "entity-rep" or "entity representation" is a term of art and does not imply the conceptual entity is modelling anything but non-intentional and non-semantic cognitive structure.

[15]The conceptualization of "that" is a valuation. Of course, this requires that "that" has a bona fide conceptualization. The issue of the meaning of "that" will be discussed in section 4.4.

this reshaping as a composition of a non-processual relationship conceptual entity, and a thing or relationship conceptual entity, that results in a modified thing or relationship conceptual entity. Thus, within the scene structure, we need only to store things and processes without separately storing non-processual relationships, so long as the processing of the S11 architecture will conduct the composition as described here.

Multiple entities can be simultaneously stored within the scene structure. Recall that in S11, a conceptual entity is modelled as an entity-rep, which is an instance of a sequence of neural activations in the scene structure. Multiple conceptual entities would be modelled as multiple entity-reps simultaneously instantiated in the scene structure as multiple sequences of neural activations, meaning that at least n sequences of neural units are required to store n entity-reps simultaneously[16]. To organize all the sequences of neural units that exist in the scene structure, the scene structure will be partitioned into *plates*. Each *plate* will contain three sequences of neural units. Each sequence in the plate will be called a *cell* (see figure 4.1). Each cell may store one of three kinds of entities, namely either a valuation, thing, or process. The cells are ordered (and thus labelled 1 to 3 in the figure) so that the entity stored by cell n can be characterized phenomenologically as having a higher focal prominence than that of cell $n + 1$. Each plate is thus a structured set, rather than a structureless bag, of cells. In practice, cell 1 will contain a valuation, and cells 2 and 3 will each contain either a thing or a process.

A single plate will thus be able to model the storing[17] of the conceptualization of the situation that is intended to be associated with such simple sentences as, e.g. "Earth moves". That is done by placing into the plate the entity-rep modelling the thing EARTH[18] for "Earth" into cell 2, and the entity-rep modelling the process MOVES for "moves" into cell 3; cell 1 would contain the entity-rep modelling the valuation START for denoting the beginning or root of this conceptualization. Similarly, the conceptualization of such simple

[16]We are thus using a form of *temporal synchrony*, i.e. treating neural units that fire in a temporally synchronized fashion as being conjoined in some sense for the purpose of representation, encoding, processing, etc.

[17]When context is clear, we will abbreviate and say that a plate stores a conceptualization, when it is actually *modelling* the storage, however it is actually implemented in the neurobiology (but hopefully, our connectionist model is oriented towards increasing neurobiological plausibility). We will use a similar abbreviation when speaking of the scene structure "storing" a conceptualization.

[18]A word or phrase written in uppercase represents the conceptualization of the situation that is intended to be associated with that word or phrase. Thus, a word or phrase written in uppercase represents a conceptual cognitive structure, and *not* a classical symbol.

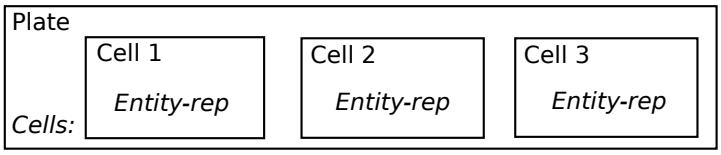

Figure 4.1: A single plate containing three cells each containing an entity-rep. The entity modelled by the entity-rep stored by cell n is characterized phenomenologically as having a higher focal prominence than that in cell $n + 1$.

clauses as, e.g. "that John ate" or "that ate John" can be straightforwardly stored as the plate containing, in sequence for cells 1 to 3, the entity-reps THAT, JOHN, and ATE, for the first example, or THAT, ATE, and JOHN, for the second example.

The many plates present will together constitute a scene structure, but a scene structure is not just a structureless set of plates. The plates will stand in certain relationships to the cells in other plates. In particular, there will exist a different plate that will be specifically assigned to each cell in a given plate (in practice, we will only allow a certain number of plates to have this property, since we have only finite computational resources) such that the assignment is *fixed*: if plate p_i is assigned to cell c_j (from another plate p_k), then: (1) the neural units $u_1, \ldots, u_m$ implementing p_i will always be used to implement p_i; (2) the neural units $u'_1, \ldots, u'_n$ implementing c_j will always be used to implement c_j; and (3) $u_1, \ldots, u_m$ will be assigned to relate always with $u'_1, \ldots, u'_n$ such that with the nature of the relationship will be shaped or characterized by the valuation entity-rep[19] stored in p_i.

The inter-plate relationship assignments allow, e.g. the thing or process entity-rep stored by a cell in one plate to stand in relation to the conceptualization stored by a second plate, such that the inter-plate relationship (i.e. the path) is shaped by the valuation entity-rep stored in the second plate (see figure 4.2). In the S11 architecture developed in this thesis, the inter-plate relation to be allowed will be that of the trajector/path/landmark (TPL) variety. Constraining the inter-plate relationships modelled to only the TPL ones also means that no plate will be assigned to the cell storing a valuation entity-rep in a plate (namely, in practice, cell 1), since valuations are the path and not the trajector in TPL relationships between plates. Thus, the plates in the scene structure (ignoring the cells in each plate)

[19]We abbreviate "the entity-rep modelling the *X-type* conceptual entity" as "the *X-type* entity-rep".

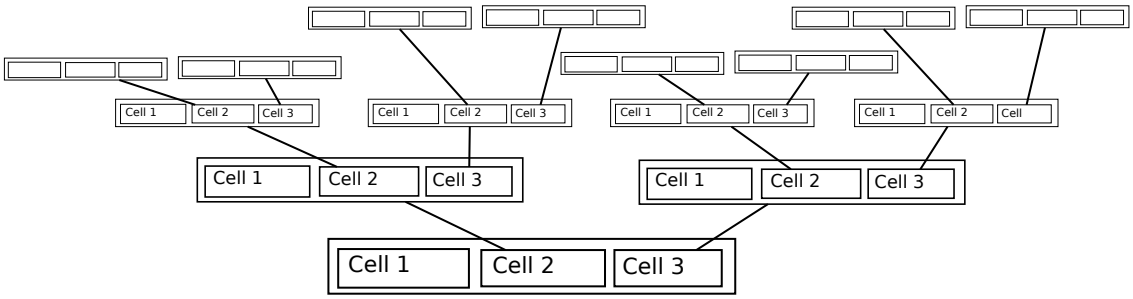

Figure 4.2: A scene structure capable of storing 45 thing, process, or valuation entity-reps, and providing a structure to interrelate them. An instance of a scene structure is the training target for a given sentence. In each scene structure, there is a distinguished plate called the *root plate*, and is illustrated in this figure as the bottom-most plate. Note the illustration's size of each plate is not significant.

form a *binary tree* structure (in the graph theoretic sense): i.e. each plate has either zero or two "child" plates, and one "parent" plate.

It is interesting to note that the scene structure is similar[20] to the *treelet* structure Marcus (2001, pps. 108-112) developed. A quick comparison may prove helpful for the reader: A *treelet* is a "preorganized, hierarchical arrangement of *register sets*. Each register set consists of an ordered set of registers that is analogous to an ordered set of bits that make up a computer's byte" (ibid., p. 108). "A given register set can hold the encoding for a variety of simple elements, such as the encoding for *cat* or *dog*..." (ibid.), where the encodings themselves could be "purely arbitrary and chosen entirely at random" (ibid.) (and thus could also be the feature based encoding used in S11). A complex structure of simple elements can be encoded into a treelet by setting the registers in each register set in the treelet to the required values, where the structural relationships between simple elements are represented by the preorganized, hierarchical arrangement of the register sets. Since the complexity of structures of simple elements will vary depending, e.g. on the sentence being comprehended, thus Marcus envisions two possibilities for representing complex structures: either (1) "the size of a given treelet might vary dynamically, depending on the state of a set of preexisting pointers that are attached only to immediately adjacent register sets"

[20]Thanks to Hadley for pointing out the similarity. The scene structure was developed independently of the treelet structure developed by Marcus (2001, pps. 108-112).

(ibid., p. 112), or (2) the representation uses "several fixed-length treelets, united by some sort of encoding system" (ibid.).

In the terminology used by Marcus, the S11 scene structure, *when every plate in it is storing at least one entity-rep*, is a single fixed-size treelet that has a binary tree hierarchical structure; each plate in the scene structure is a register set, and each cell in a plate is a fixed subset of registers in the register set; and each register is a neural unit. Since the scene structure is a fixed-size treelet (when every plate contains at least one entity-rep), there is an upper-bound to the complexity of structures of simple elements that can be stored. If some plates in the scene structure are empty (i.e. storing no entity-reps), then only those plates that are not empty (i.e. is storing at least one entity-rep in its cells) are considered part of the treelet represented in the scene structure (i.e. at most one treelet can be represented in the scene structure in S11 at any one time, but the treelet is re-sized dynamically using the large supply of pre-existing plates in the scene structure). In this manner, any arbitrary scene structure can be implemented in terms of treelets, and thus scene structures can be alternatively described using the treelet notation; but of course, not every treelet can be implemented as a scene structure.

Returning now to our previous discussion on the inter-plate relationships, it may seem that constraining the inter-plate relationships to only the TPL kind is restrictive, but it is in fact sufficient for quite an interesting range of conceptualizations, including ones that are described by sentences with multiple levels of grammatical embedding. For example, consider *sentence A* as follows: "Alice said that Bill believes that Cindy claims that Doris swallowed a spider". In our scene structure, sentence A requires only five plates to model its conceptualization (see figure 4.3).

Sentence A has also been formally analyzed by Langacker (2008, p. 417) using Cognitive Grammar, and it is analogous with his analysis that we may conclude from figure 4.3 that the conceptualization of the situation that is intended to be associated with sentence A "is then a nonhierarchical, chain-like grammatical structure, where each clause is linked to the next by a correspondence. Semantically, this link consists in the expectation that the schematic landmark will be further specified" (ibid., p. 418). Of course, our analysis, as in figure 4.3, requires that "that" has a bona fide conceptualization; the issue of the meaning of "that" will be discussed in section 4.4.

Notice that the scene structure displays two properties that we may call *conceptual systematicity* and *conceptual compositionality*. The scene structure is *conceptually systematic*

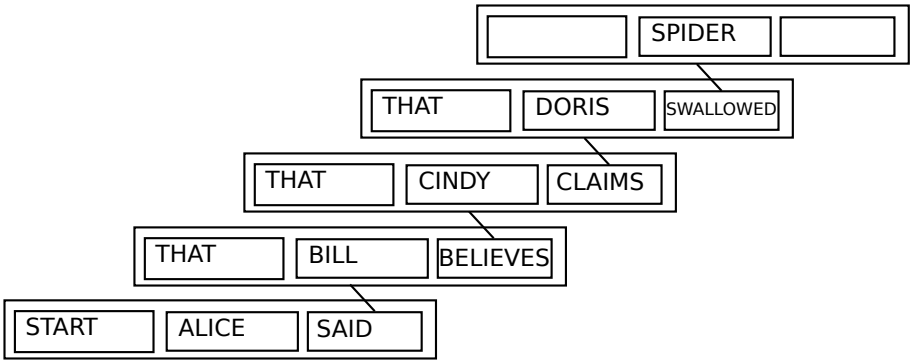

Figure 4.3: An instance of a scene structure that models the conceptualization of the situation that is intended to be associated with sentence A. For clarity in the diagram, all the empty plates from the scene structure, i.e. those with no cell containing neural activations, are not shown in this diagram even though they are still present in the scene structure.

in that it is capable of storing the conceptualization of "John loves Mary" just in case it is capable of storing the conceptualization of "Mary loves John". This is a different kind of systematicity than semantic systematicity, since the latter refers to an ability to *assign meaning to words* (recall the definition from section 2.4), whereas the former refers to an ability to *store a conceptualization of a situation* of the external (physical) world.

The scene structure is *conceptually compositional* in that when it stores the conceptualization of "John loves Mary", it tokens the cognitive structures of the conceptual entities JOHN, LOVES, and MARY[21] in such a way that each can be found and be isolated from the rest. That is, if a neural unit is involved in storing one conceptual entity, then it will not be involved in storing any other conceptual entities simultaneously. This kind of compositionality is, of course, roughly similar to the *concatenative compositionality* that is used in Classical symbol processing systems (see van Gelder, 1990; Aizawa, 1997a, p. 50).

That conceptual compositionality is similar to symbolic concatenative compositionality does not, however, make this architecture a Classical one, or make the scene structure a kind of Fodorian Language of Thought (LoT). As Fodor and Pylyshyn (1988) made it quite clear, Classical theories postulate a Language of Thought or mental representation where

[21]Note it is the cognitive structures, and *not* classical symbols, that are tokened.

"the semantic content of a (molecular) [mental] representation is a function of the semantic contents of its syntactic parts, together with its constituent structure" (ibid., p. 8). Thus, in Classical theories, the "expressions in [Fodorian] LoT are *mental representations*, and they *represent* something: entities in the world. Put differently, Fodor insists that LoT is *intentional*" (Jackendoff, 2002, p. 279), it has "semantic properties (i.e. are meaningful, can be interpreted as being about something)" (van Gelder, 1990, p. 366), and it has *semantic content*.

Notice how drastically different the Classical view is from what is being modelled in this thesis. Here, an English sentence utterance generate phonological cognitive structure that gets associated with a conceptual cognitive structure (modelled in S11 as a scene structure). Conceptual cognitive structure does not have meaning, does not have semantics, does not have semantic content, and is not intentional; however, conceptual cognitive structure (a spatial-temporal arrangement of atomic cognitive elements, which we assumed are neural activations in the brain, that can also be characterized phenomenologically) *is* the semantics, or put more specifically, the *cognitive significance*, for English sentences *as utterances* that generate certain phonological cognitive structures[22].

The brain will either (1) implement multiple areas of storage such that each can be modelled by a scene structure, or (2) implement multiple areas of storage such that each can be modelled by a plate, and such that an underlying neurobiological processing mechanism exists that can be modelled as the numerous plates being reorganized into multiple scene structures. In either case, each scene structure would model the storage of just one construal or aspect of the conceptualization of a complete situation, such that the situation may be conceptualized from currently occurring perceptual stimulations or be conceptualized from what would phenomenologically be described as vivid imagination. This complete conceptualization is referred to as the *conceptual substrate* by Langacker (2008, p. 54). This thesis only deals with a single scene structure and does not model the entire conceptual substrate, a task best left to future research.

[22]This echos the analysis by Jackendoff, who argued that: "Semantic/conceptual structure does not *have* a semantics, it *is* the semantics for language" (Jackendoff, 2002, p. 278-9). The major difference of this thesis from Jackendoff's proposal is that the S11 architecture does *not* have an explicit generative linguistic syntax processor or module, whereas his proposal for an architecture does. It may be objected that symbolic Markerese, from Classical architectures, also *is* the semantics for language, but the real issue here is that symbols in Markerese requires its own *separate* (secondary and auxiliary) interpretive semantics to enable the Markerese symbols to *represent* and be *about* things in the external world. Conceptual cognitive structure does not have such a separate intentional semantics *at all*. See section 4.1

Recall that process relationships involve an arrangement of the relationships and entities involved over a *span* of time. The modelling of a process as a simple sequence of neural activations in this model makes this an incomplete model, as the constituent features are fully manifested, static, and atemporal. In future work, a dynamic and real-time model would have to be developed, but the atemporality in this model should not bother us too much now, as none of the neural network systems proposed, and none of the systems surveyed in section 2.3, take into account embodied cognitive structures that are dynamically manifested over processing time.

4.4 Connectionist Modelling of Conceptual Entities

Recall that each conceptual entity will be modelled by an entity-rep, which is a sequence of neural activations, and that entity-reps will be encoded based on features, so that if an entity-rep has three features, then the sequence of neural activations encoding that entity-rep will have activations only in the three neural units encoding those three features. In particular, each conceptual entity is formed by a subset of features, and the full list of features may be partitioned into these five separate categories[23]:

1. intrinsic thing properties (e.g. red, heavy, tall),
2. extrinsic thing properties (e.g. location relative to some conceptual ground),
3. intrinsic process properties (e.g. rapid, smooth, movement that has occurred),
4. extrinsic process properties (e.g. location or movement relating things in primary and secondary focal prominence),
5. valuation properties (e.g. entities' level of focal prominence, role in the trajector/path/landmark cognitive schema, degree of objectivity).

Thus, thing conceptual entities (what are ordinarily expressed by nouns), and any non-processual relationship conceptual entity that has as its trajector a thing conceptual entity

[23]The examples provided here of each category are only meant to be incomplete models of whatever properties the processing of perceptual stimulations would come up with in the human brain in an unsupervised manner. It is not meant that the brain would literally come up with the "heavy" property from seeing heavy things.

(what is ordinarily expressed by adjectives), will overlap in the various thing-properties. Further, process relationship conceptual entities (what are ordinarily expressed by verbs), and any non-processual relationship conceptual entity that has as its trajector a relationship conceptual entity (what is ordinarily expressed by adverbs), will overlap in the various processual-properties. Finally, the conceptual entities ordinarily expressed by subordinators (i.e. "that", "while", and "with") and the START conceptual entity will overlap in the valuation properties.

The reason conceptual entities ordinarily expressed by subordinators are given valuation features is that, especially for some subordinators, they are traditionally viewed as purely syntactic, but in Cognitive Grammar, they are viewed as meaningful and thus the problem is rather "to say just what their meanings are" (Langacker, 2008, p. 413). Since at least some of the subordinators are commonly viewed as purely syntactic, this suggests that their meanings do not, on a traditional realist view, directly contain *disembodied* external referents, and thus suggests that their meanings are, at least in part, valuations within the Cognitive Grammar framework. For example, Langacker suggests that the meaning of "that"[24] is that "it explicitly marks the proposition expressed as an **object of conception**...[and] as being construed objectively rather than subjectively"[25] (Langacker, 2008, p. 444).

A conceptual entity consisting of features resulting from the processing of perceptual stimulations is modelled as an entity-rep encoded by a sequence of neural activations (recall section 4.3), such that each neural activation in the sequence would correspond with a particular feature. Note that since each plate consists of three cells, and each cell is used to store a entity-rep, then each plate will contain three entity-reps (one per cell).

As an example of how a conceptual entity is modelled by an entity-rep, consider the entity CHAIR. Let e(CHAIR) be the vector representation of the entity-rep sequence of

[24]In saying that the conceptualization of "that" is a valuation, we are not claiming or rejecting the possibility that the conceptualization may involve a relation that exists objectively and independently of a mind (but note that does not mean a knower can infallibly know any such relation). We are simply stating that the conceptual features are not extracted from the perception (e.g. through ocular irradiation) of a whole or part of a physical entity, possibly through time, but instead may simply map to relationships or processes previously abstracted from relationships or processes internal to the speaker's *body* that were prompted during past experiences.

[25]The use of "proposition" here is meant to refer to the *conceptualization* for the clause whose meaning is traditionally said to be a proposition. Recall that in Cognitive Grammar, meaning is imagistic and experiential, and not propositional in the style of a Language of Thought (Langacker, 2008, p. 32).

neural activations for CHAIR. Let the features for CHAIR be, e.g. the following: :physical, :short, and :heavy[26]. Let e(:x) be the vector representation of the entity-rep sequence of neural activations representing feature :x. Then e(CHAIR) = e(:physical) + e(:short) + e(:heavy), where the vector summation is performed element-wise such that the sum of two elements is bounded between zero and one (i.e. the minimum and maximum allowable neural activation levels for neural units in the scene structure).

As another example, consider the entity RED-CHAIR, which is the conceptualization expressed by "red chair". CHAIR will be, for the purpose of this example, as analyzed in the previous example. Let the features of RED be, e.g. :red (note the difference between RED and :red is that the former is a conceptualization, and the latter is a perceptual feature possibly contained in that conceptualization). Then we have e(RED-CHAIR) = e(RED) + e(CHAIR) = e(:red) + e(:physical) + e(:short) + e(:heavy). A similar analysis would apply to such conceptualizations as RUN and SWIFTLY-RUN, etc.

In this thesis, each feature is assigned a random and unique 1-of-K vector representation as its entity-rep, where $K = 106$ (since there are 106 features to be considered). Thus, the entity-rep for each conceptual entity would also be encoded as a vector of length 106 (of course, an entity composed of n features, for $n > 1$, would *not* be encoded as a 1-of-K vector but would instead be encoded as a n-of-K vector).

For a listing of the features actually assigned to each entity-rep to be considered in this thesis, see the tables in appendix A. The feature assignments (i.e. features assigned to each entity-rep) listed in appendix A are in part inspired by the feature assignments used by Hadley et al. (2001), and also in part inspired by the features of conceptualizations suggested by Jackendoff (2002, p. 312-3) and Langacker (2008, p. 72-3, 260). Of course, since each feature is assigned a random and unique vector representation as its entity-rep anyway, the list of possible features used does not matter; what does matter is the *overlap* in features of the various conceptual entities modelled, which is specified in appendix A. The feature assignments used in this thesis, though in part motivated by conceptual semantic theories, must be treated as a simplification used in the modelling for the purpose of building a proof of concept system. Having said that, as mentioned above, it is assumed that these features that make up conceptual entities are capable of being causally created in the human brain through the neurobiological processing of perceptions and stimulations of a situation — a

[26] We will use the colon-prefixed notation, i.e. ":x", to mean that x is a feature.

process that has been studied in connectionist research, particularly by those employing Deep Belief Nets (e.g. Bengio, 2007, p. 2-3), which can in fact learn a higher level neuro-encoding of perceptual features in an entirely unsupervised manner.

4.5 Connectionist System Architecture: Overview

Recall that Cognitive Grammar proposes a conceptual semantics that is based foundationally on a pairing of conceptual and phonological cognitive structures for each piece of language in the mind of a language user. The conceptual cognitive structure is said to be in the *semantic pole* of the pairing, and the phonological cognitive structure in the *phonological pole*. The whole "language processing stack" (from words up to sentences) is formed from these pairs, or the recombination of these pairs into new pairs through various kinds of compositionality (Langacker, 2008, pps. 15-8).

Consider a case like "jar" and "lid", to form "jar lid". The phonological word "lid" is on one pole of the phonological/semantic pair lid/LID[27], where LID is the meaning (recall from section 2.6 that in conceptual semantics, meaning resides in conceptualization). Similarly, we can also have the pair jar/JAR.

To compose the two phonological/semantic pairs requires two processors, one for the phonological pole, and one for the semantic pole. The phonological processor can just concatenate the two phonological words together into "jar lid" — since the connectionist architecture being proposed in this thesis is not meant to produce language (just to comprehend it), this processor is not of interest here.

The semantic processor, on the other hand, has to combine the JAR and LID conceptualizations together. Simple concatenation does not suffice except in the simplest of cases. For instance, "little" and "green men" do not compose simply to the expected result of MARTIAN-ALIEN. A similar story can be told for verb phrases, e.g. involving symbols run/RUN and quickly/QUICKLY. Furthermore, the semantic processor will have to be able to compose the semantic pole of more complex phonological/semantic pairs like those for "The cat", "is on", and "the mat", into the correct conceptualization.

[27] From here on, a symbol shown with both phonological and semantic poles will be written in the form *x/Y*, where *x*, written as the word or phrase in lowercase, represents the phonological cognitive structure of the word or phrase, and where *Y*, written as the word or phrase in uppercase, represents the conceptual cognitive structure of the situation that is intended to be associated with the word or phrase.

For this reason, we will have to create in S11 a sophisticated mechanism that can produce the correct scene structure given a complete sentence to process, in order to model the semantic processor. This processing will be called *inference*, which will be described in the rest of this chapter, and the S11 architecture being developed in this thesis will be a hybrid[28] type of connectionist architecture. If successful, S11 will *infer* the scene structure appropriate for the given sentence, which will be processed phonologically word by word. The *learning* process whereby S11 will learn how to make such inferences will be described in section 4.7.

Given a sentence, e.g. "Earth moves", the S11 system will process it one word at a time from left to right. Each word will be encoded phonologically, although we will only use an *extremely* simplified model of phonology as we are not in particular interested in a study of phonology in this thesis. Thus, from hereon, all mention to phonology is to the simplified model to be described presently.

In particular, a phonological word is a 5-tuple sequence. Each item in the sequence is a syllable form. Each syllable form is a 26-tuple sequence, such that each item in it encodes an alphabet character. So for now, we assume each phonological word is at most five syllables long. Finally, in encoding a word into a phonological word, the word is partitioned into sub-strings delimited by vowels so that each sub-string is encoded as a single syllable form, and thus the sequence of syllable forms form a single phonological word encoding that original word. A word encoded phonologically will be called a *word-rep*. This encoding is done using ordinary programmatic means, as opposed to using a connectionist network (again, because phonology is not our focus here).

So given a sentence, the connectionist system S11 will in fact be processing a sequence of word-reps (which models the required phonological cognitive structure). At the start of processing a sentence, the scene structure is blank and contains only neural unit activation values of zero. The system progressively builds up to the correct scene structure of the sentence as it processes each word-rep in sequence. The S11 hybrid connectionist system is functionally modular, with the following modules:

1. Lexicon Module (FFN)
2. Plate Chooser (FFN)

[28] Recall the definition of hybrid architectures from section 2.7.

3. Plate Focuser (CSP)
4. Plate Indicator
5. Auxiliary Plate Indicator
6. Cell Chooser (FFN)
7. Cell Indicator
8. Entity Compositor (FFN)
9. Scene Structure
10. Goal Object Resolver (CSP)
11. Information Channels (CSP)

The modules labelled "FFN" are traditional feedforward neural networks. The modules labelled "CSP" are implemented in the software as classical symbol processing functions, and they are the reason why S11 is a *hybrid* connectionist architecture rather than a *purely* connectionistic one. The Plate Focuser, however, can be implemented in a connectionistic architecture in a manner to be described in section 5.3. Further, both the Plate Focuser and the Goal Object Resolver cognitive modules will be argued as oriented towards increasing biological plausibility in chapter 5. A system of interconnections amongst the modules is required, of course, to channel information from one module to another; such a system is called the *Information Channels*, and it will also be argued as oriented towards increasing neurobiological plausibility in section 5.4.

The Plate Indicator, Auxiliary Plate Indicator, Cell Indicator, and the Scene Structure are storage areas for sequences of neural activations, and may be implemented as cyclically firing neurons (similar to the conjunctive binding nodes used by Hadley & Cardei, 1999, and Hadley & Hayward, 1997) that use a linear activation function, and that simply cyclically fire with an activation value equal to the activation it last received; they may also be *purely* functional modules that do not have an independent neurobiological existence, for example, as in the case of the Cell Indicator, which could be implemented as simply the neurons of the output-layer of the Cell Chooser. For the purpose of this thesis, we will assume they are cyclically firing neural units, similar to binding nodes, in order to simplify the implementation and description of the functioning of all the modules.

4.6 Connectionist Processing — Inference

Let us now focus on the details of how the connectionist system S11 will handle the processing of a single word-rep. The inference process involves all the previously mentioned modules *except the Goal Object Resolver and Auxiliary Plate Indicator*, and can be briefly summarized as taking the following steps (in order):

1. If the given word-rep is the start of a sentence, then the Scene Structure and the Cell Indicator are reset so that they all have neural unit activation values of zero, and furthermore, the Plate Indicator is reset so that the root plate becomes the plate in focus

2. Lexicon Module processes the word-rep as input

3. Plate Chooser uses various neural unit activation values from within the system as input to infer its output-layer

4. Plate Focuser uses various neural unit activation values from within the system as input to infer the activation values to be stored in the Plate Indicator

5. Cell Chooser uses various neural unit activation values from within the system as input to infer its output-layer, which is subsequently subjected to a winner-take-all process the result of which is stored in the Cell Indicator

6. Entity Compositor uses various neural unit activation values from within the system as input to infer an entity-rep

7. The entity-rep is stored in the cell indicated by the Cell Indicator (i.e. the *cell in focus*) from the plate indicated by the Plate Indicator (i.e. the *plate in focus*)

8. Repeat from step 1 using the next word-rep in the given sentence

The above steps are a brief outline. Details of the individual modules are presented in the rest of this chapter. A more complete summary of the steps will be presented afterwards in section 4.6.7.

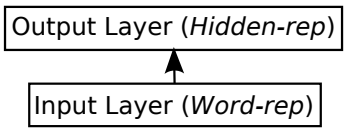

Figure 4.4: Lexicon Module for mapping word-rep to hidden-rep, using a 1-layer perceptron. Arrows show full-connectivity.

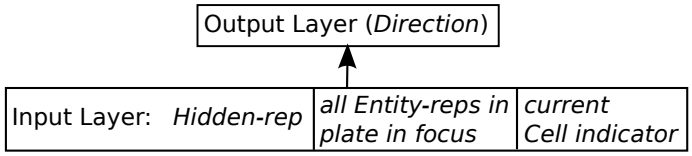

Figure 4.5: Plate Chooser for selecting which plate, from the scene structure, to bring into focus, using a 1-layer perceptron. Arrows show full-connectivity.

4.6.1 Lexicon Module

The purpose of the *Lexicon Module* is to "translate" each word-rep to be processed into a distributed representation internal to the system that allows the system to infer the correct entity-rep, and to infer which cell is the correct one to store the entity-rep. The distributed representation is learned from the data, and is called the *hidden-rep*. Note that the hidden-rep is *not* a form of an entity-rep, so despite the word "Lexicon" in the name, the Lexicon Module is *not* intended to model the association of a conceptual entity or a conceptualization to a phonological cognitive structure.

The Lexicon Module is an ordinary one-layer[29] feedforward neural network (also known as a perceptron) that takes as input a word-rep, and that will infer from the input a hidden-rep (see figure 4.4). The number of output-layer neural units in the Lexicon Module will be two-thirds the number of input layer neural units. Like all other modules, the Lexicon Module will use the error-backpropagation algorithm for connection weight tuning. The output-layer neural units in the Lexicon Module will use a hyperbolic-tangent activation function, since the inferred hidden-rep will be passed on to other neural network modules for further processing. Thus, the hidden-rep is a distributed representation of the input that the Lexicon Module learns to produce for each word-rep.

4.6.2 Plate Chooser and Plate Indicator

The purpose of the *Plate Chooser* is to, on the basis of the hidden-rep and some other information, and in conjunction with the Plate Focuser module, help choose the correct

[29] I will adopt the terminology recommended by Bishop (2007, p. 229) for feedforward neural networks, where a *N-layer neural network* has N layers of adaptive weights, since "it is the number of layers of adaptive weights that is important for determining the network properties" (ibid.).

plate to become the new plate in focus. The Plate Chooser does this by inferring the "direction" of the new plate in focus relative to the previous plate in focus (i.e. relative to the plate indicated by the Plate Indicator).

The *Plate Indicator* is a sequence of P neural units, where P is the number of plates that exists in the scene structure. Each neural unit corresponds to one plate in the scene structure, such that the neural unit with the highest activation value identifies the plate that is considered to be *in focus*. In practice, there will be only one neural unit with an activation value of one, and all other units will have an activation value of zero. There exists also an *Auxiliary Plate Indicator*, which is just another sequence of P neural units with the same properties as the Plate Indicator just described, and serves as a copy of the Plate Indicator whenever the Plate Indicator is updated — the Auxiliary Plate Indicator is used only by the Plate Focuser and *only during learning*, and so it will not be discussed again until section 4.7.2.

The reason for designing the Plate Chooser and Plate Indicator in this manner can be understood by considering the metaphorical intuition[30] behind the task the Plate Chooser accomplishes. Specifically, there is always a single plate in the scene structure that is considered in focus, and which plate is in focus changes depending on the word currently being processed and on which previous plate was in focus (and the entity-reps stored in that plate). Thus, the Plate Chooser is metaphorically navigating or moving the "in focus label" around the scene structure from one plate to the next. Hence, the Plate Indicator will always have one neural unit with an activation value of one (and all others zero) to indicate which plate is currently in focus; further, the Plate Chooser is thus tasked to output a "direction" that allows S11 to appropriately modify the Plate Indicator to indicate the next plate in focus, and thus effectively move the "in focus label" around the scene structure. There is a deeper metaphorical intuition behind why we would want the Plate Chooser to help navigate or move the "in focus label" around the scene structure, an intuition based on how cognition operates according to an analysis by Cognitive Grammar theory, but an explication of that intuition is best left for section 4.6.3, after the operation of the Plate Focuser is explained.

Let us return to the connectionistic implementation details of the Plate Chooser. The

[30]By "metaphorical intuition", I mean an imagistic or phenomenological understanding that, though it may not be completely technically correct, provides a "what it is like"-ness to improve our comprehension of what may otherwise be very dry or be lacking in motivation.

Plate Chooser will have an output-layer of five neural units, each of which we may interpret as a kind of "direction"[31] in the scene structure relative to the plate currently in focus. The directions are, namely: *root*, *self*, *parent*, *left-child*, and *right-child*. Specifically, consider the scene structure illustrative diagram in figure 4.2, and consider only the plates (i.e. ignore the cells in each plate): it clearly has a binary tree structure, where each plate has either zero or two "child" plates, and one "parent" plate — except for one plate that has no parent (the one at the base of the tree). The plate with no parent plate is called the *root plate*, and it is the plate indicated by the neural unit representing the *root* "direction". The other directions are to be interpreted relative to a given plate, namely, the plate in focus: the *self* plate relative to the given plate is, obviously, the given plate; the plate referred to by the other directions relative to a given plate should be equally obvious now within the binary tree structure of the scene structure. How the S11 system manages to set the Plate Indicator to the correct neural activation values, given the previous plate in focus and the direction inferred by the Plate Chooser, will be explained when the Plate Focuser is described, as the Plate Focuser is what manages this functionality.

The Plate Chooser is an ordinary one-layer feedforward neural network (see figure 4.5). It will use the error-backpropagation algorithm for connection weight tuning. The output-layer neural units will use a *soft-max* activation function

$$z_i = \frac{e^{a_i}}{\sum_{j \in \{\text{output neurons}\}} e^{a_j}}$$

where a_i is the total activation received by output neural unit i, and z_i is the activation value of output neural unit i. The soft-max activation function is chosen as it is favored for the task of K-class classification (Hastie, Tibshirani, & Friedman, 2009, p. 383), which is what the Plate Chooser is essentially doing when it is choosing one of five "directions" as described above. The Plate Chooser will take as input:

1. the hidden-rep inferred by the Lexicon Module,

2. the entity-reps stored in every cell in the plate in focus in the scene structure, and

[31] These are not pointers in the traditional computing science sense (or at least not in the sense used in the C programming language), but can be interpreted to be labels for nodes in a binary tree (a tree in the graph theory sense) such that which node carries which label is always relative to the single node indicated as currently active.

3. the current Cell Indicator[32].

These three sequences of neural activations will be concatenated together into one single input-layer of activations.

4.6.3 Plate Focuser

The purpose of the Plate Focuser during inference is to take as input the Plate Indicator, which shows the previous plate in focus, and the Plate Chooser's output-layer activation values, which shows the inferred "direction", to output the new Plate Indicator. Just as feedforward neural networks (FFNs) that employ error-backpropagation for learning operate differently during learning than during inference, the Plate Focuser, which is not a FFN that uses error-backpropagation for learning, also operates differently during learning than during inference. This section will describe how the Plate Focuser operates during inference.

The function of the Plate Focuser during inference is of course easy to implement using traditional programmatic methods, and they are implemented in that way in S11 (which is one of the reasons why S11 is a *hybrid* connectionist architecture rather than a *purely* connectionistic one). They can, however, in principle be implemented purely connectionistically in a manner to be described in section 5.3. The way in which they are implemented using traditional programmatic methods will be described presently.

During inference, the Plate Chooser's output-layer is first subjected to a winner-take-all process in order to decisively choose one "direction". This is done by letting the neural unit with the strongest level of neural activation "win" and take on the maximum level of activation (here defined to be 1), while all other units will take on the minimum level of activation (here defined to be 0). The winner-take-all process is used commonly within connectionist architectures (e.g. Hadley et al., 2001; Hadley & Cardei, 1999; Rumelhart & Zipser, 1986). Next, the Plate Indicator and the Plate Chooser's output-layer after the winner-take-all process are both treated as separate 1-of-K vectors. Finally, ordinary programming easily allows us to produce a 1-of-K vector to be stored in the Plate Indicator to indicate the new plate in focus based on the two input vectors.

[32] Actually, it is the sequence of activation values stored by the Cell Indicator that is used as input: not the neural nodes referred to by the term "Cell Indicator". What is being referred to should be clear from usage. The same usage is true of the term "Plate Indicator" as well.

The reason for designing the Plate Focuser in this manner can be understood by considering the metaphorical intuition behind the task the Plate Focuser accomplishes. Specifically, the Plate Focuser is essentially just a module to support the Plate Chooser in selecting one plate to focus upon, given another plate was previously focused upon. Given the "directions" that that Plate Chooser can select from, the possible plates to focus upon are just the root plate and the plates surrounding the given previous plate in focus. Recalling that each plate can encode the meaning of such simple sentences as "Jon laughed", and can encode the meaning of such simple clauses as "that chased Sally", metaphorically, each plate can be interpreted as encoding a tiny portion of the complete conceptualized situation, i.e. the conceptual substrate. This semantic interpretation is further strengthened by the semantic relationship that exists between neighbouring plates.

Recall the semantic relationship that exists between neighbouring plates: each cell in each plate is associated to a different plate, so that the association allows, e.g. the thing or process stored by a cell in one plate to stand in relation to the conceptualization stored by a second plate. The inter-plate relation defines the path in the trajector/path/landmark relationship between the entity stored in the involved cell, and the conceptualization stored in the involved plate.

Appealing to our metaphorical intuition, we may thus imagine the scene structure as a space of meaning fragments, and the Plate Chooser and Plate Focuser work together to identify the salient fragment on the basis of the previous salient fragment. Put differently, we may imagine the Plate Chooser and Plate Focuser as working together to identify the salient event within the situation, given the previously salient event. By doing so, the Plate Chooser identifies the (hopefully) correct event that the word, currently being processed by the system, is "about". Once the correct event, i.e. the correct "location" in the space of meaning fragments, has been identified, the system can perform further processing on the identified fragment independently and in isolation from all other fragments.

The sequential choosing and focusing that occurs between the Plate Chooser and Plate Focuser is, in fact, an instance of the *sequential scanning* cognitive phenomenon that is employed, e.g. for "mentally tracking an event as it unfolds through time, that is, scanning sequentially through it along the temporal axis" (Langacker, 2008, p. 111); it is also employed in apprehending the nested locative construction, e.g. as appears in the sentence, "Your camera is upstairs, in the bedroom, in the closet, on the shelf" (ibid., p. 195), where one must mentally locate the trajector (the camera) by "successively 'zooming in' to smaller

and smaller areas" (ibid.). Thus, the existence of the Plate Focuser and Plate Chooser within our connectionistic model is supported in a "top-down" manner, by appealing to theories of conceptual semantics and Cognitive Grammar.

4.6.4 Information Channels

The *Information Channels* module of S11 is not a processing module that transforms information. Instead, it is a term introduced to refer to all the copying of neural activation values that is necessary between all the other modules. For instance, the hidden-rep inferred by the Lexicon Module is copied into the input-layer of all the other three feedforward network modules. As discussed above, the Plate Chooser also requires the entity-rep stored in every cell in the plate in focus in the scene structure to be copied into its input-layer. All this copying is performed in the software implementation of S11 by ordinary programmatic methods (e.g. copying the floating point numbers representing neural activations from one array to another), and in part because of this, the implemented model is a hybrid connectionist model.

The Information Channels module will be argued as being oriented towards increasing neurobiological plausibility in section 5.4.

4.6.5 Cell Chooser and Cell Indicator

The purpose of the *Cell Chooser* is to, on the basis of the hidden-rep and some other information, choose the correct cell to become the new cell in focus (within whichever plate is currently in focus). The Cell Chooser does this by outputting neural activation values through the three neural units contained in the output-layer in the Cell Chooser, such that each unit corresponds to one of the three cells in a plate.

The Cell Chooser is an ordinary one-layer feedforward neural network (see figure 4.6). It will use the error-backpropagation algorithm for connection weight tuning. The output-layer neural units will use a soft-max activation function as it is favored for the task of K-class classification, which is what the Cell Chooser is doing when it is choosing one of the three cells as described above. The output-layer consists of three neural units, as mentioned, corresponding to the three cells in any arbitrary plate in the scene structure. The Cell Chooser will take as input:

1. the hidden-rep inferred by the Lexicon Module,

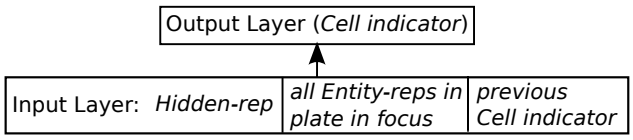

Figure 4.6: Cell Chooser selects the cell to bring into focus in the plate currently in focus, using a 1-layer perceptron.

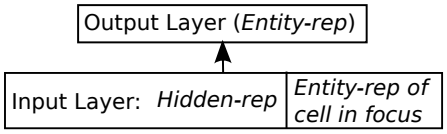

Figure 4.7: Entity Compositor for composing together the entity-rep stored in the cell in focus in the plate in focus, with the hidden-rep inferred by the Lexicon Module, using a 1-layer perceptron.

2. the entity-reps stored in every cell in the plate in focus in the scene structure, and

3. the current Cell Indicator.

These three sequences of neural activations will be concatenated together into one single input-layer of activations.

Once the Cell Chooser infers the output-layer activation values, they will be copied to the *Cell Indicator*, which is a sequence of three neural units. The neural units in the Cell Indicator would then be subjected to a winner-take-all process, whereby the neural unit with the strongest level of neural activation value will "win" and take on the maximum level of activation (here defined to be 1), while all other units will take on the minimum level of activation (here defined to be 0). The winning neural unit identifies the cell that is considered to be in focus, within the plate that is currently in focus as indicated by the Plate Indicator.

To further the metaphorical intuitive image, first described in section 4.6.3, that we had of the scene structure as modelling a space of meaning fragments, or a structure of interrelated conceptualizations of events making up a conceptualization of a situation, let us also look briefly at the metaphorical intuition behind the Cell Chooser. Since each plate, as mentioned, can store a meaning fragment, or a conceptualization of an event from a conceptualization of a situation, the Cell Chooser is focusing in on one aspect of a meaning fragment, or of a conceptualization of an event, when it identifies a cell as being in focus (an "aspect" being one of the three conceptual entities modelled and stored as an entity-rep in the plate in focus). Specifically, it identifies the (hopefully) correct "location" for an entity-rep (that is modelling a conceptual entity within the event conceptualization modelled and stored by the plate in focus), so that either the appropriate entity-rep can be "placed" into

that location to replace the entity-rep already there, or that the entity-rep in that location can be changed appropriately, by using the Entity Compositor to manipulate the entity-rep in the identified "location" independently and in isolation from all other meaning fragments in the scene structure, and in accordance with the word currently being processed by the system.

4.6.6 Entity Compositor

The purpose of the *Entity Compositor* is to infer, on the basis of the hidden-rep H and the entity-rep E_1 stored in the cell in focus in the plate in focus, the correct entity-rep E_2. If correctly inferred, E_2 models the conceptual entity resulting from the composition of the conceptual entity modelled by E_1, and the conceptual entity associated with the phonological cognitive structure modelled by the word-rep that the Lexicon Module used to infer H. Notice that, intuitively, the Entity Compositor can be said to be in a hidden state defined by H, when it then manipulates E_1 to transform or mutate E_1 into E_2, and it is doing so on E_1 independently and in isolation from all other entity-reps in the scene structure.

As a practical example, a cell may contain the entity-rep encoding the conceptual entity RED, and the hidden-rep may be inferred by the Lexicon Module from processing the word-rep of "apple", after having just processed the word-rep of "red". Now, suppose the Entity Compositor is processing, using as input, the entity-rep for RED and the hidden-rep for "apple". The Entity Compositor is then intended to infer the entity-rep encoding the conceptual entity RED-APPLE. The inferred entity-rep is then stored in the same cell that just previously stored the entity-rep for RED, i.e. the inferred entity-rep is "copied" into the the aforementioned cell (thereby "overwriting" whatever was previously stored in that cell). Note, though, as per our previous discussion in section 4.6.4, that such copying and transfer of information is performed by Information Channels.

The Entity Compositor is an ordinary one-layer feedforward neural network (see figure 4.7). It will use the error-backpropagation algorithm for connection weight tuning. The output-layer neural units will use a *logistic sigmoid* activation function

$$z_i = \frac{1}{1 + e^{-a_i}}$$

where a_i is the total activation received by output neural unit i, and z_i is the activation

value of output neural unit i. The logistic sigmoid activation function is chosen as it is favored for the task of inferring a vector of values between zero and one, which is what is required as the entity-rep is just such a vector.

If we wish, we may interpret each component value in the entity-rep vector as the strength to which the Entity Compositor "believes" the semantic feature encoded by that component should be present in that entity-rep. Put differently, the Entity Compositor is essentially performing a *multi-label 2-class classification*, where the presence of each semantic feature in the entity-rep is represented as two classes (namely, *present*, and *not-present*) to discriminate between, where the classification between the two classes is expressed as a probability between zero and one, and where there are multiple semantic features to label the entity-rep with.

The Entity Compositor will take as input:

1. the hidden-rep inferred by the Lexicon Module, and
2. the entity-rep stored in the cell in focus from the plate in focus from the scene structure.

These two sequences of neural activations will be concatenated together into one single input-layer of activations.

4.6.7 Summary of Inference Procedure

The beginning of section 4.6 had sketched out a simplified outline of how the entire connectionist system S11 operates. Before moving on to explaining how S11 manages to learn anything, we can provide a much more detailed outline now:

1. If the given word-rep is the start of a sentence, then the Scene Structure and the Cell Indicator are reset so that they all have neural unit activation values of zero, and furthermore, the Plate Indicator is reset so that the root plate becomes the plate in focus
2. Lexicon Module infers the hidden-rep using the word-rep as input
3. Plate Chooser infers its output-layer using the following as input: the hidden-rep, the Cell Indicator, and all entity-reps stored in the plate in focus as indicated by the Plate Indicator

4. Plate Focuser uses the Plate Indicator and the output-layer of the Plate Chooser as input to infer the activation values to store in the Plate Indicator, which is then subjected to a winner-take-all process

5. Cell Chooser infers its output-layer using the following as input: the hidden-rep, the Cell Indicator, and all entity-reps stored in the plate in focus as indicated by the Plate Indicator (note that the Plate Indicator has possibly been changed from its initial state by the previous step in the processing). The output-layer is then copied into the Cell Indicator, which is then subjected to a winner-take-all process

6. Entity Compositor infers an entity-rep using the following as input: the hidden-rep, and the entity-rep in the cell in focus in the plate in focus in the scene structure (note that the Cell Indicator has possibly been changed from its initial state by the previous step in the processing)

7. The entity-rep inferred by the Entity Compositor from the previous step is finally copied into the cell in focus in the plate in focus in the scene structure, replacing whatever entity-rep was previously stored in that cell

8. Repeat from step 1 using the next word-rep in the given sentence

4.7 Connectionist Processing — Learning

As previously described in section 4.5, the inference process allows the connectionist system to infer the scene structure appropriate for a given sentence by processing it phonologically word by word. The *learning* process (or *learning phase*), whereby the connectionist system will *learn* how to make such inferences, will be described in this section.

The connectionist system manages to learn, and remember what has been learned, through changing, from their initial values, the connection weights between neural units in these four feedforward neural network (FFN) modules: Lexicon Module, Plate Chooser, Cell Chooser, and Entity Compositor. The connection weights are, initially, uniform randomly chosen from between -0.02 and 0.02. The connection weights are then adjusted through a *training process*, which will be a form of error-backpropagation that will be described in section 4.7.4. The complete learning process involves more than merely the training of the

four FFN modules, however, since the learning process must also be capable of obtaining the *training signals* required for the training of each of the FFN modules.

Although the details of the training process will be described in section 4.7.4, an overview of the process is required to define some of the terminologies that will be necessary for understanding the higher level functioning of the learning process. A FFN module is trained using error-backpropagation by changing the connection weights between its neural units in the presence of training data. Error-backpropagation is a supervised training method, meaning that the training data must contain the complete input-output pairs (as opposed to unsupervised training, where only the inputs are required). For this reason, the input-output pairs are often called *input-target pairs*, to emphasize that the output is the desired target of the input in the given pair.

In the *batch training* version of error-backpropagation, the weight adjustments are calculated as a function of the existing connection weights and the whole batch of training data, and then the weight adjustments are applied once. Of course, the weight adjustments can be *re*calculated and applied on the basis of the new connection weights and the same batch of training data. Of course, the weight adjustments can be *re*calculated and applied... Etc. Each time the weight adjustments are *re*calculated and applied is called an *epoch*, or a "pass", through the training set data. The process of calculating and applying the weight adjustments in *each* epoch is called *weight tuning*. Multiple epochs of training (i.e. weight tuning) are usually required.

There is also a different version of error-backpropagation training that serves as an alternative to batch training called *stochastic* or *online training*. In online training, the weight adjustments are calculated as a function of the existing connection weights and just one sample out of the whole batch of training data, and then applied immediately to the connections. The weight adjustments are then recalculated and applied on the basis of the new connection weights and a different sample from the training data. The adjustments are iteratively recalculated and applied for each of the samples in the training data. A training *epoch*, or a "pass" through the training set data, has occurred once each sample of the training data has been used once in the calculation and application of weight adjustments. Multiple epochs of training are usually required, and the weights are tuned, per epoch, for as many times as there are samples in the training data. For various reasons, it is advantageous to choose online training over batch training (for details, see Bishop, 2007; LeCun, Bottou, Orr, & Müller, 1998), and we will use online training here for our S11 architecture.

4.7.1 Training Data and Goal Object Resolver — Cognitive Foundations

As explained in section 4.1, we presume a language learner associates the perceptual stimulations of an utterance (i.e. utterance stimulations) to the perceptual stimulations of a situation (i.e. situation stimulations). Thus, language users learn language by building up a dictionary of pairs of cognitive structures generated neurobiologically from sentence utterance stimulations, and by situation stimulations (i.e. stimulations of either the whole situation or some fragment of it. Recall footnote 9 in section 2.6 regarding the distinction between "whole" versus "fragment"). Furthermore, the learning process requires the language learner to build up a sufficiently large training data set in order to work out the semantically significant component stimulations (e.g. word stimulations) of the sentence utterance stimulations in the language, associate them to conceptual entities from conceptualizations of situations, and then use such associations to work out the conceptualizations of novel sentences (so long as all the semantically significant parts of the novel sentences have been seen in the data set).

Of course, a language learner does not go around the physical world picking cognitive structures out of thin air. Instead, since the required cognitive structures are generated neurobiologically from stimulations causally created by physical processes, in effect, the language learner simply needs to be exposed to many sentence utterances spoken by many language natives in many situations. In section 4.1, this part of the language learning process was referred to as the collection of a large data set of ASM pairs: pairings between affirmative stimulus meanings and sentence utterances.

During the process of building up the training data set of ASM pairs, the language learner would presumably observe a native speaker utter a sentence under a certain situation, and then the learner associates that sentence utterance stimulations to that situation stimulations (recall that the association is assumed to be valid on the presumption of the Principal of Charity). All of this has been said before in section 4.1. What we should carefully analyze now is how underdetermination by observation can be mitigated. For example, if a native speaker says "the cat chased the dog" in the presence of a cat *and* a dog, how can a learner figure out that the "cat" utterance means the cat stimulations rather than the dog stimulations in that sentence? Given that it is impossible to observe a cat in isolation, i.e. in the absence of anything else observable (i.e. to observe a cat literally in a void), no amount of observation can, in principle, determine exactly that "cat" means the

cat stimulations, and that it is *intended* to be associated with the cat presumed to exist and causing those stimulations.

For the practical purpose of language learning, this issue turns out to be less problematic than initial appearance would suggest. In the course of language learning, a learner is often taught by a teacher who would point out various physical things in the situation that the learner and the teacher are both observing. For example, the teacher may, repeatedly, point at a cupboard and say "cupboard" simultaneously — but note that though the teacher *intends* to point at the cupboard, all the teacher can do is point in a direction determined by ocular stimulations so that it appears *to the teacher* that he or she is pointing at the cupboard, and then the teacher can only *hope* that the ocular stimulations the student receives will be sufficient for the student to work out what the teacher intended to point at.

As another example, the teacher may point at the tail of a dog while saying "the tail", then say "of" without pointing, and finally point out the entire dog while saying "the dog"[33] — but note once again that the teacher can only *intend* to point at a thing by pointing in a direction based on all the perceptual stimulations the teacher has received, and then all the teacher can do is hope the student receives sufficient perceptual stimulations for the student to work out what the teacher intended to point at.

The "pointing" referred to in the above two examples may be literal, as when the teacher uses their hand or finger to point, or may come as a set of behavioural cues. No amount of pointing, however, will solve the principled problem of underdetermination, since the act of pointing by the teacher is itself just observable data (i.e. ocular stimulations) for the learner. In practice, on the other hand, pointing is an invaluable source of data.

Pointing is an invaluable source of data because it helps *constrain* what thing conceptual entity in a conceptualized situation a learner should focus on at a given time. The physical thing pointed to by a teacher is called a *goal object*. In S11, the process that models the cognitive process of identifying and focusing onto the conceptualization of the goal object is called the *entity focusing process*.

The ability for animals to identify a goal object is well known in psychological research due to Pfungst's study into what is now known as the *Clever Hans effect* (Pfungst, 1911).

[33] I suspect the reader has experienced or seen this kind of teaching done with children. I personally recall having experienced the first example, and I have personally taught English to a child in a similar manner as the second example. These are certainly not idle examples, and similar examples can be easily found in schools and homes.

This effect was first documented in the twentieth century, when a horse named "Hans" was believed to be capable of simple arithmetic and other mental tasks, until the psychologist Pfungst showed that Hans was actually recognizing and responding to small unintentional behavioural cues (e.g. posture, facial cues) of his trainer and audience (ibid.). For example, by observing the way the trainer stood in relation to a series of cloths, Hans was able to approach a selected cloth from the series (see ibid., pps. 21, 81, 156)[34].

The Clever Hans effect has been observed with other animals as well, including dogs (Lit, Schweitzer, & Oberbauer, 2011), and these studies show just how powerful an indicator of goal objects that behavioural cues can be. The Clever Hans effect shows that the pointing done by a teacher are, in fact, often behavioural cues other than literally pointing with a finger, and need not be linguistically based at all. In fact, children "[a]fter the first few years of life... can effortlessly follow the gaze of their adult teachers [and] share attention with teachers to common goal objects during social interactions" (Grossberg & Vladusich, 2010, p. 940). Grossberg and Vladusich (2010) describes the burgeoning field of behavioural, neurophysiological, and neuroimaging research on this phenomenon, and produced a neural model of how the brain can produce such behaviours as shared attention of goal objects.

For the purpose of this thesis, we will assume a cognitive module called the *Goal Object Resolver* (GOR) will be able to correctly identify the conceptualization of the goal object the learner ought to attend to, based on the behavioural cues provided by the native speakers *without* appealing to linguistic data. Put differently, during the entity focusing process of the learning phase, the GOR will enable S11 to identify and focus on the correct entity-rep modelling the conceptualization of the goal object. The GOR is a biologically inspired abstraction or simplification of what clearly must exist in the brain, as shown from the evidence discussed above regarding shared attention of goal objects, but the GOR module as proposed here is *not* literally meant to be found in the brain, because a complete model of shared attention of goal objects would be beyond the scope of this thesis (this thesis is focused on language, not on behavioural cues).

In fact, the GOR module will be implemented classically by common programmatic means in this thesis, without any attempts to implement it connectionistically. The reason

[34]Hans often made errors in the selection, unless the selected cloth was at the extreme left or right of the series. It turns out errors were made because as the horse approached the cloths, he lost sight of the questioner, and was then unable to react appropriately to the behavioural cues provided. This problem rarely occurred when the selected cloth is at the ends of the series as cues indicating the extreme ends was a lot simpler and more robust for the horse to follow. See Pfungst, 1911, p. 81.

the GOR is implemented classically is that something like the GOR module must exist in the brain, and thus must be implementable in a biologically realistic manner anyway (a more detailed discussion of these issues will be presented in section 5.1): in a thesis focused on language learning, there is no reason to commit to a connectionist framework of how perceptual stimulations generated by behavioural cues might be neurobiologically processed, since we already know that (1) a similar module must exist in the brain to support the functionality displayed by humans, horses, dogs, etc., of acquiring shared attention of goal objects, (2) it can operate independently of language learning (as demonstrated in horses, dogs, etc.), and (3) a neural model of acquiring shared attention of goal objects already exist elsewhere (see Grossberg & Vladusich, 2010). Due to the fact that the GOR module is implemented classically, the S11 connectionist architecture is a hybrid one.

All in all, the training data that the connectionist system will be given will be a set of N pairs (X_i, T_i), for $i \in \{1, \ldots, N\}$, where X_i is the sequence of word-reps encoding the sequence of words of a given sentence, and where $T_i = (T_{i1}, T_{i2})$ such that T_{i1} is the target scene structure's entity-reps, and T_{i2} is the behavioural cues stimulations from the teacher that allows the GOR to select the correct entity-reps (modelling conceptualizations of goal objects) out of the target scene structure. Notice that the scene structure is not somehow synchronized with the word order in the sentence represented, and so the word order is learned from the training data set based on behavioural cues stimulations provided by the teacher to support the shared attention of goal objects.

The specific constitution of the training data set will be described in chapter 6. Note that since the GOR is a simplification, the set of behavioural cues stimulations T_{i2} will not be literally presented here. Instead, we simply assume that the required behavioural cues stimulations are provided when appropriate, that they are interpreted correctly by the GOR, and so the correct goal object conceptualization will be identified at the appropriate times by the GOR during training. What is specifically identified during training is, for each word-rep in the training sentence, the particular entity-rep of the conceptual entity the word is "about" in the scene. e.g. in "Jon moved red chair", Jon is identified as the goal object for "Jon", the moving motion for "moved", and the red chair is *separately* and repeatedly identified for each word "red" and "chair". Note that the GOR is *not* providing grammatical information about each word, as evidenced by the GOR resolving the goal object for both "red" and "chair" as the *same* red chair conceptual entity.

This simplification of the process in animals, which produces shared attention of goal

objects, into the GOR module may appear as an appeal to a *deus ex machina* device. It must be emphasized, however, that such a cognitive module simply *must* exist, and be able to work out the conceptualization of the goal object using non-linguistic behavioural input (e.g. bodily attitude, gaze), based on the cognitive research on horses, dogs, and infants, as described above (a more detailed discussion of these issues will be presented in section 5.1). Furthermore, it will turn out that a bootstrapping process can in fact be employed during the learning phase's entity focusing process, so that the GOR need not be used all the time during learning, thereby modelling a more cognitively realistic process of gradual bootstrapped learning.

It may also appear that the entity-reps T_{i1} would only be available if a lot of prior language learning had already taken place, but recall that the conceptualization formation process is intended to be possible through a non-linguistic and unsupervised method as described in section 4.1. In particular, recall that the capabilities of Deep Belief Net architectures were discussed to show that there is already existing research that is suggestive of how something like the scene structure (modelling conceptualizations) might be formed experientially, without intentionality, and in a possibly neurobiologically plausible manner (see also section 5.1).

4.7.2 Plate Focuser

Recall our discussion of the Plate Focuser's operation during inference from section 4.6.3, wherein it was stated that the operation of the Plate Focuser during learning is different than during inference. This section will describe how the Plate Focuser operates during learning.

During learning, the Plate Focuser will take as input the previous Plate Indicator (recall, from section 4.6.2, that it would have been copied to the Auxiliary Plate Indicator), and the current Plate Indicator (as presently set by the Goal Object Resolver), to output the direction indicator that must have been inferred by the Plate Chooser in order to get to the current Plate Indicator from the previous Plate Indicator during inference. The purpose of this operation during learning is so that the Plate Focuser can find the correct direction indicator as the training signal for the Plate Chooser to train with using error-backpropagation.

The function of the Plate Focuser during learning is of course easy to implement using

traditional programmatic methods, and they are implemented in that way in S11 (which is one of the reasons why S11 is a *hybrid* connectionist architecture rather than a *purely* connectionistic one). They can, however, in principle be implemented purely connectionistically in a manner to be described in section 5.3. The way in which they are implemented using traditional programmatic methods will be described presently.

Given the Auxiliary Plate Indicator and the Plate Indicator, the former indicating the previous plate in focus, and the latter representing the current plate in focus (as determined by the Goal Object Resolver), we can simply operate the Plate Focuser *as though we are doing inference* (recall from section 4.6.3 how it operated during inference) on each of the available "directions" to determine, by brute-force[35], which is the correct direction that would have taken us from the previous plate in focus to the current plate in focus. Lastly, output the correct 1-of-K vector representing that direction.

4.7.3 Training Signals and Learning

We can now describe how the connectionist system S11 will learn when given just one single training pair (X_i, T_i) from the training data, with $T_i = (T_{i1}, T_{i2})$ as in section 4.7.1. Also, let X_i be a sequence of m word-reps: $X_{i1}, X_{i2}, \ldots, X_{im}$. First, the entity-reps of the cells in the scene structure are *clamped* to that of T_{i1}. Next, the system performs a *forward-pass* using X_{i1} as the input word-rep. The forward-pass is a process mostly similar to the process of inference as described in section 4.6.7. Next, *training signals* are provided to the Entity Compositor, Plate Chooser, and Cell Chooser feedforward network (FFN) modules. Finally, all four FFN modules in the system perform error-backpropagation. The specifics of these three processes will be described in this section below.

Clamping of the scene structure is the process by which the scene structure in S11 is set to a given T_{i1} target scene structure, such that the scene structure in S11 can no longer be changed by the system. This is a standard process assumed to occur in all kinds of connectionist systems in order that weight tuning may occur. The *forward-pass* process is essentially the inference process applied to a single word-rep, *except* that the entity-rep inferred by the Entity Compositor is *never* copied into the cell in focus in the plate in focus in the scene structure (refer to section 4.6.7 for a summary of the inference process, but be

[35]Obviously, brute-force is computationally wasteful, and a more efficient program can be easily written, but for expositional purpose, this is a sufficient description here.

reminded to omit step 7 where the inferred entity-rep is copied).

After the forward-pass is completed based on the input of X_{i1}, the S11 system will provide training signals to clamp the output layers of the Entity Compositor, Plate Chooser, and Cell Chooser FFN modules with. The *training signal* for these FFN modules is just the desired target output-layer activation values. By using the term *clamping* of a FFN module's output layer, what we mean is that the desired output values are stored in the output neural units of that module for the purpose of comparison with calculated neural activation values in order to carry out the error-backpropagation algorithm (see section 4.7.4 for details).

Once a FFN module's output layer has been clamped to a training signal, that module can then perform the error-backpropagation algorithm (see section 4.7.4). The provision of the training signals and the performance of the error-backpropagation algorithm on all four FFN modules in the system will together constitute the *backward-pass* process of the system. The forward-pass and backward-pass process, together and performed in that order, constitute the *learning* process.

For the purpose of describing how the training signals are derived, let us suppose the Goal Object Resolver (GOR) has indicated cell j from the scene structure as containing the entity-rep to focus on (modelling the conceptualization generated by the goal object). We will further assume that the manner in which the GOR is able to indicate cell j to focus on is by setting the Cell Indicator and Plate Indicator appropriately. The three training signals required will then be created in the manner to be described presently.

For the Entity Compositor, the training signal is just the entity-rep E_j from cell j. Thus, the system will simply copy E_j to the Entity Compositor, wherein the output-layer will be clamped to E_j so that error-backpropagation can be performed on the Entity Compositor. It will be argued in section 5.4 that such a system to copy E_j can be implemented in a connectionistic manner through the use of Information Channels.

For the Cell Chooser, the training signal is just the Cell Indicator. The Cell Chooser's output-layer is simply clamped to the activation values in the Cell Indicator, and then error-backpropagation is performed on the Cell Chooser.

For the Plate Chooser, recall from section 4.7.2 our discussion regarding how the Plate Focuser operates during learning: the Plate Focuser will take the Plate Indicator as input, and provide as output a sequence of neural unit activation values representing the "direction" that would be desirable for the Plate Chooser to have inferred. That Plate Focuser

output is thus the required Plate Chooser training signal, and is copied to the Plate Chooser, wherein the output-layer will be clamped to it. The Plate Chooser will then be in a position to perform error-backpropagation.

As will be apparent from our discussion in section 4.7.4 on the FFN error-backpropagation algorithm used, once the Entity Compositor, Cell Chooser, and Plate Chooser have performed error-backpropagation, we will be able to "extract" the "error" of the hidden-rep that had been copied to those three module's input-layer during the forward-pass (namely, by computing δ_j of equation 4.3). The "error" of the hidden-rep can then be summed *element-wise*, and the resulting sequence of neural unit "error" values will then be the training error signal for the Lexicon Module. The Lexicon Module does *not* use this sequence of "error" values to clamp its output-layer, but instead uses the sequence of "error" values to continue performing the error-backpropagation algorithm (i.e. using those "error" values as the δ_j of the neural units of the Lexicon Module's output-layer).

Another way of describing how the Lexicon Module performs error-backpropagation is to say that the relevant "errors" will be backpropagated from the other three FFN modules into the Lexicon Module. A simple way of seeing how this might work within a connectionistic system is to imagine that the neural units from the Lexicon Module's output-layer are actually shared amongst the input-layers of the other three FFN modules. In this way, the four FFN modules can be said to be "fused" together in that they share corresponding neural units where appropriate. Running the error-backpropagation algorithm with the other three FFN module's output-layer clamped as described above will naturally backpropagate the errors back into the Lexicon Module, and will tune the Lexicon Module's connection weights accordingly.

Once all four FFN modules have performed error-backpropagation, the backward-pass phase is complete, and all four FFN modules would then have had their neural unit connection weights tuned once. The S11 system will then repeat the forward-pass and backward-pass procedure with the next word-rep, namely X_{i2}, then again with X_{i3}, X_{i4}, ..., up to and including X_{im}. Once the system has performed the forward and backward-passes for each word-rep in the given sentence in the training pair (X_i, T_i) as described in this section, the system can then go on to repeat the learning procedure described in this section with the next training pair: (X_{i+1}, T_{i+1}). An epoch of learning is constituted by applying the learning procedure, as described in this section for one training pair, on each of the training pairs in the training data set. Multiple epochs are often required.

4.7.4 Error-Backpropagation in a Feedforward Network

The particular error-backpropagation (also called backpropagation or BP) algorithm to be used on the four feedforward neural (FFN) networks in the S11 system will be described in this section. Recall from section 2.2 that a FFN network consists of a connected directed acyclic graph of neural units. Each unit z_j applies a nonlinear activation function h onto the activation a that the unit receives, where a is a linear combination of the unit's input signals $x_1, \ldots, x_D$, to arrive at an output signal (called its activation value). That is, if given D units connecting into unit j, the activation value of unit j is z_j:

$$z_j = h(a_j) \tag{4.1}$$

$$a_j = \sum_{i=1}^{D} w_{ji} x_i + w_{j0} \tag{4.2}$$

We refer to parameters w_{j0} as the bias and w_{ji} as the connection weight on the connection from neural unit i to unit j. The nonlinear activation function is often chosen to be a sigmoidal function, e.g. the logistic sigmoid or the hyperbolic tangent functions. The activation function is sometimes also called the *transfer function*. Thus, to avoid confusion with the output signal (activation value) of a neural unit, we may also call the activation a neural unit receives its *pre-transfer value*.

During training, an input is given to the FFN along with its desired output, and the connection weights in the FFN are adjusted by BP operating on the input/output pair. Of course, many such training pairs can be used for training, and BP is applied to each training pair separately in online training. The following description is for the application of BP on a single training pair.

There are three phases to BP when it is applied to a single training pair: *forward propagation*, *backward propagation*, and *gradient-descent*. They will be described in turn, starting with forward propagation.

Of the set of neural units, the *input units* are those that have outgoing connections to, but no incoming connections from, other neural units. Since the FFN is a connected directed acyclic graph of units, once the input units have an activation value set (e.g. set to the values of the input pattern), neural activations may propagate "outwards" iteratively to neural units that have incoming connections only from input units or other neural units that have activation values already calculated by equation 4.1, until all non-input neural

units have activation values so calculated. This iterative process of propagating neural activations is called the *forward propagation* phase of training an FFN.

During supervised training of the FFN, forward propagation is performed first, and then neural activations of the output neural units calculated during forward propagation are compared to the clamped values of the output neural units (i.e. the desired neural output activations). By using the term *clamping*, what we mean is that the desired output values are stored in the output neural units. The purpose of clamping the output neural units is so that when the neural activations of the output units are calculated by equation 4.1 during forward propagation, the desired and calculated values can then be compared for the purpose of the *backward propagation* process. For each output neural unit, the comparison is made by a simple subtraction of the desired activation value from the calculated activation value, and the resultant difference is called the *error signal* of that output neural unit.

Backward propagation is performed after the forward propagation is complete. For simplicity, let us define the *error network* as the FFN but with the direction of all connections between neural units reversed. Thus, e.g. the output neural units will have no incoming connections in the error network. Backward propagation begins with the error signal of the output units calculated as described above (i.e. the subtraction of the desired activation value from the calculated value). Next, the error signal is iteratively propagated "outwards" in the *error network* in a manner similar to how activation values are propagated during forward propagation in the FFN. The key difference is that the equation used to calculate a non-output neural unit's error signal is in fact:

$$\delta_j = h'(a_j) \sum_{k \subset K} w_{kj} \delta_k \tag{4.3}$$

wherein h' is the derivative of the activation function, a_j is the pre-transfer value of neural unit j, and the sum is taken over the set K of neural units that are connected to unit j in the *error network*. Once every neural unit has its error signal calculated, the backward propagation of errors is complete.

Finally, the optimization algorithm of *gradient-descent* is performed whereby each neural unit connection weight is adjusted by a small *training rate* value. Specifically, the connection

weight from neural unit j to unit i in the FFN is adjusted by

$$-\eta \Delta w_{ij} \tag{4.4}$$

$$\Delta w_{ij} = \delta_i z_j \tag{4.5}$$

That is, the adjustment is by the negative of the product of the error signal of unit i, the activation value of unit j, and the training rate η. The actual weight adjustment rule used, however, is slightly modified, in order to speed up training, and to avoid letting the network learn models that do not generalize well. Specifically, the weight update is:

$$w_{ij}^{\text{new}} = (1-\gamma) w_{ij}^{\text{current}} - \eta \Delta w_{ij} + \rho (w_{ij}^{\text{current}} - w_{ij}^{\text{previous}}) \tag{4.6}$$

where γ is the (connection) *weight decay rate*, and ρ is the *momentum rate*. In this thesis, we set $\rho = 0.8\varepsilon$, $\gamma = 0.0008\varepsilon$, and $\eta = 0.2\varepsilon$. The parameter ε is set to 0.02 for the Cell Chooser, Entity Compositor, and the Lexicon Module, and set to 0.03 for the Plate Chooser.

Adding the *momentum* term into equation 4.6 helps the network train more quickly by providing each update with some information from past updates. Momentum provides a "damping" effect that lessens oscillations where the error surface is irregular and where the gradient changes sign regularly, and speeds up the weight changes where the error surface is long and flat. For greater details on the use of momentum, refer to Moreira and Fiesler (1995).

Weight decay is a *regularization* technique that penalizes weights with magnitude that grew too big. Regularization helps ensure the FFN learns a model that generalizes better to novel input. The use of weight decay is justified by realizing that the weight adjustment of equation 4.4 is actually the result of taking the partial derivative of the sum-of-squares error function $\sum_k (y_k - t_k)^2$ with respect to the weight to be updated, and the weight decay factor in the update rule of equation 4.6 results from taking the same partial derivative but with the error function regularized as $\sum_k (y_k - t_k)^2 + \sum_l w_l^2$, where k ranges over the training samples available, and l ranges over *all* connection weights in the FFN. For details of weight-decay and regularization, refer to Hastie et al. (2009, p. 398).

Chapter 5

Neurobiological Issues

5.1 Preliminaries

In section 2.8.1, it was argued that connectionist architectures like that of Hadley and Hayward (1997) may be rightly criticized by a variant form of the argument initially put forth by Aizawa (1997a, 1997b). Recall from section 2.8.1 that the variant form of Aizawa's criticism points out that a connectionist architecture does not have a very satisfying level of scientific explanatory power (to explain systematicity in humans) when it is missing the *specific contextual details* that would justify the architectural decisions made as resulting in an architecture that is scientifically explainable or justifiable for the purpose of explaining systematicity in humans — "specific contextual details" is emphasized because as explained in section 2.8.1, merely saying "evolution made it so" is not very satisfying[1]. It was further noted that one way to provide the sufficient contextual details would be to ensure the connectionist architecture is supported by other cognitive or neurobiological theories, that is to lean on results from other cognitive or neuroscientific theories, and then let *those* theories be explained by others through some evolutionary account[2].

In chapter 4, the theories of Cognitive Grammar (one form of cognitive linguistics) was

[1]Recall from section 2.8.1, however, that Hadley only claimed to offer an explanation of systematicity "in some *possible* cognitive agent" (R. F. Hadley, personal communication, 24 June 2011) that might not necessarily be human (ibid.), and that might be existing in a possibly *different* world (R. F. Hadley, personal communication, 30 June 2011), whereas this thesis is interested in furthering an understanding of a *specific* species of cognitive agents here on Earth, namely, humans (*Homo sapiens*).

[2]Recall that this idea of *sharing* the responsibility of explaining phenomena is not revolutionary and is well defended. See footnote 25 from section 2.8.1 for details.

appealed to in support of the model implemented as the S11 architecture in this thesis. The support from Cognitive Grammar lets us say that, to the degree it is so supported, the S11 architecture is oriented towards increasing psychological plausibility.

Although it must be emphasized that S11 is a *hybrid* connectionist architecture rather than a *purely* connectionistic one, it would be preferable if we have some indication that the S11 architecture is not entirely out of the realm of neurobiological possibility on the proviso we understand that it is difficult to achieve neurobiological realism in any connectionist model in the first place. If that much can be shown, then it can be said that the S11 architecture is also oriented towards increasing neurobiological plausibility. Therefore, *this chapter is intended only to provide* further *theoretical support of the model implemented in S11*; put differently, issues of neurobiological plausibility are just one of many pillars of support on which the S11 architecture rest upon. Notably, this chapter is included in this thesis mainly to support, in advance, a reply to the variant form of Aizawa's criticism from section 2.8.1: that is to show that S11 provides, or begins a research program to provide, a bona fide scientific explanation of systematicity in humans, rather than simply exhibiting some interesting properties. All we need at this point, therefore, are broad strokes to show an orientation towards increasing neurobiological plausibility in S11, rather than fine details to shown neurobiological realism.

Much of this chapter is necessarily speculative, as the goal is to show a possibility that may justify the claim that S11 is oriented towards increasing neurobiological plausibility. On the whole, the question of how exactly any connectionist architecture may be concretely implemented in neurobiology is simply an open research question.

The S11 architecture consists of many modules, and we will briefly examine each in turn in the following. The Plate Focuser and Information Channels will require more involved discussions and will be discussed in separate sections below.

Recall that the Plate Indicator, Cell Indicator, and the Scene Structure are storage areas for sequences of neural activations. Recall that they may be implemented as cyclically firing neurons (similar to the conjunctive binding nodes used by Hadley & Cardei, 1999, and Hadley & Hayward, 1997) that use a linear activation function, and that simply cyclically fire with an activation value equal to the activation it last received, but that they may also be *purely* functional modules that do not have an independent neurobiological existence, for example, as in the case of the Cell Indicator, which could be implemented as simply the neurons of the output-layer of the Cell Chooser. For the purpose of this thesis, recall

we assumed that they are cyclically firing neural units, similar to binding nodes, in order to simplify the implementation and description of the functioning of those modules. Given that the neural units in those three modules are so similar to binding nodes (insofar as they are cyclically firing neural units), it seems their use in S11 makes S11 no less oriented towards increasing neurobiological plausibility as compared to their use in the architectures proposed by Hadley and Cardei (1999), Hadley and Hayward (1997), etc.

Next, consider the Goal Object Resolver (GOR). Recall from section 4.7.1 that for the purpose of this thesis, we assumed the GOR is a cognitive module that will be able to correctly identify the conceptualization generated neurobiologically by (e.g. from ocular irradiation) the goal object the learner ought to attend to, based on the behavioural cues (received by the learner, e.g. as ocular irradiation) provided by the native speakers *without* appealing to linguistic data. The GOR is a biologically inspired *abstraction, simplification, or first approximation* of what clearly must exist in the brain, as shown from the evidence discussed in section 4.7.1 regarding shared attention of goal objects, but the GOR module as proposed in this thesis is *not* literally meant to be found in the brain, because a complete model of shared attention of goal objects would be beyond the scope of this thesis (this thesis is focused on language, not on behavioural cues). Therefore, the GOR module is implemented classically by common programmatic means in this thesis, without any attempts to implement it connectionistically.

It may be argued that the GOR is a contentious model of the neurobiological origin for the behaviour of acquiring shared attention of goal objects (a behaviour that is undeniably present in many species, as shown from the evidence discussed in section 4.7.1), especially in regards to the level of accuracy and detail at which the GOR operates[3]. It should be emphasized, however, that the GOR module in S11 is just a biologically inspired *abstraction, simplification, or first approximation*, and especially the behaviour of *bootstrapping* in learning is omitted in the model entirely — admittedly, this is a simplification in the model, but as frequently occurs in cognitive modeling, choices have been made to explore certain complexities while simplifying others.

For example, the complete accuracy built into the GOR to model the goal object identification process, and thus the complete accuracy built into the entity focusing process during learning (since in S11, the entity focusing process during learning relies completely

[3]Thanks, especially, to R. F. Hadley (personal communication, 24 June 2011) for pointing out this issue.

on the GOR), is surely an idealization of what does occur in real life during child learning. I would suggest that in real child learning, a more gradual bootstrapping process occurs whereby prior learning of grammar and vocabulary, having made use of the neural mechanism for acquiring shared attention of goal objects (a neural model of which has been proposed by Grossberg & Vladusich, 2010), will have a tremendous positive effect on later learning in terms of what is selected as the goal object when a word is presented in a sentence. In fact, already, there are "models that simultaneously learn to segment words from phoneme strings and learn the referents of some of those words" (Johnson, Demuth, Frank, & Jones, 2010, p. 1018) where it can be shown "that there is a synergistic interaction in the acquisition of these two kinds of linguistic information" (ibid.). Although the issue of bootstrapping or interactive learning is an issue for future research, preliminary experiments on a modification of S11 has already shown the promise of bootstrapped learning (see section 6.3).

It must be reemphasized that the reason the GOR is implemented *classically* is that something *like* the GOR module must exist in the brain, and thus must be implementable in a neurobiologically realistic manner anyway: in a thesis focused on language learning, there is no reason to commit to a connectionist framework of how perceptual stimulations generated by behavioural cues might be neurobiologically processed, since we already know that (1) a similar module must exist in the brain to support the behaviour displayed by humans, horses, dogs, etc., of acquiring shared attention of goal objects, (2) it can operate independently of language learning (as demonstrated in horses, dogs, etc.), and (3) a neural model of acquiring shared attention of goal objects already exist elsewhere (Grossberg & Vladusich, 2010). Based on these reasons, the GOR module should not be seen as to impede, in broad strokes, the claim that S11 is oriented towards increasing neurobiological plausibility.

The Lexicon Module, Plate Chooser, Cell Chooser, and Entity Compositor modules are feedforward networks (FFN) that employ error-backpropagation learning. They will be discussed together in section 5.2. The Plate Focuser will be discussed in section 5.3, and Information Channels in section 5.4

5.2 Error-Backpropagation

It should be noted that it has been "widely believed that Hebbian learning is probably closer to biological reality than the commonly used method of backpropagation of error" (Hadley & Hayward, 1997, p. 5). Therefore, by employing error-backpropagation in this thesis, the S11 architecture may *seem* to be farther from neurobiological plausibility than, e.g. the architecture developed by Hadley and Hayward (1997). The situation is, however, not so clear cut.

As described in chapter 4, all the feedforward network (FFN) modules of S11 employ only one layer of adaptive weights (i.e. there are *no hidden layers* of neural units). Movellan (1991) has shown that *contrastive Hebbian learning* is *equivalent* to error-backpropagation for networks with no hidden neural units. Thus, although the FFN modules all employ error-backpropagation in this thesis for convenience of modelling, it has been shown that they can in fact be converted into connectionist networks that use Hebbian style learning. Specifically, each FFN module would need to be converted into what is known as a *continuous Hopfield model*.

There is, furthermore, other work that shows error-backpropagation in a FFN, even with multiple hidden layers of neural units, can be approximated by Hebbian style learning (Xie & Seung, 2003). The work of Xie and Seung, of course, represents a second possible option for converting the FFN modules in S11 into networks that explicitly use Hebbian style learning.

It was mentioned in section 4.7.3 that the four FFN modules in S11 could be thought of as being "fused" together to form a single FFN. In that case the resulting FFN would have one hidden layer of neural units and the result from Movellan would not apply. In this case, of course the result of (Xie & Seung, 2003) would still apply.

It may be argued[4], however, that the works of Xie and Seung (2003), and Movellan (1991), do not show the Hebbian style learning employed use neurobiologically plausible connection weight change increments. Put differently, some may argue that the increments used for connection weight adjustments in the Hebbian learning may not, *prima facie*, be neurobiologically plausible *if, e.g. the connection weights in a connectionist architecture are interpreted to model the strength of the synapses between dendrites and axons from*

[4]Thanks, especially, to R. F. Hadley (personal communication, 24 June 2011) for pointing out this issue.

adjacent neurons, since the weight adjustments in the connectionist model may be, e.g. any real number[5] between -1 and 1, while biological synapses "have rather course granularity for synaptic weight settings (about 10 levels of possible weights for a typical neuron)" (R. F. Hadley, personal communication, 13 June 2011). While that may be true, it must be viewed in the context of the type of abstraction, over neurobiological reality, used in connectionist research like this thesis and also like those surveyed in sections 2.3 and 2.7.

Specifically, in connectionist research, the architectures proposed are not necessarily intended to model at the *single biological neuron level of description* (i.e. each neural node in a connectionist model does not necessarily model a single biological neuron[6]). "Admittedly (as [Hadley] initially stressed), the nodes and links in our model must be regarded as abstractions which emerge from lower-level structures and processes." (Hadley, 1997, p. 574). Thus, it may not be possible to *directly* compare the connection weights of the links between neural nodes in connectionist architectures with biological synapses in terms of the granularity of synaptic weight settings.

For example, if each connectionist node is in fact modelling a group of 100 biological neurons (such that each neuron in a group modelled by a node has at least one synapse connecting it to a neuron from an adjacent group of neurons modelled as an adjacent node), the links between nodes may then be modelling a group of synapses numbering somewhere between 100 and 10000. In such an example, if the connection weight of the links between nodes model, e.g. the average synaptic strength of the underlying synapses being modelled, and if the synaptic strength is measured on a scale between zero and one (such that there are only ten levels of possible synaptic strengths, and thus the change in synaptic strength is only as small as approximately 0.11), the allowable increments, in terms of connection weight adjustments of the connectionist links, can in fact be as small as approximately 0.0011 (for a total of about 910 link weight settings, if a link models 100 synapses, since $0.11/100 = 0.0011$) to 0.000011 (for a total of about 90910 link weight settings, if a link models 10000 synapses).

Clearly, when connectionist nodes are interpreted to model groups of biological neurons

[5]When implemented on a modern computer, the actual allowable increments will, of course, be dictated by the type of floating point numbers used in the programming language of choice.

[6]In fact, if we wish to model at the single biological neuron level of description, our connectionist architecture would likely have to be implemented in terms of *leaky integrate-and-fire* models of *spiking neurons* (Stewart & Eliasmith, 2009).

rather than single neurons[7], connection weight change increments used in Hebbian style learning in the connectionist model can in fact be much smaller than it may, *prima facie*, appear. Therefore, in the context of connectionist research, wherein we are motivated to produce architectures inspired by neurobiology without necessarily achieving full neurobiological realism, the use of Hebbian learning with small weight increments (that do not *prima facie* appear neurobiologically plausible) does not actually impede a claim to an orientation towards increasing neurobiological plausibility so long as we do not claim (and this thesis does not claim) to model at the single biological neuron level of description.

As a result, the employment of error-backpropagation in this thesis cannot, solely on its own, be used to argue[8] that the S11 architecture is any farther from neurobiological plausibility than many other connectionist architectures that use Hebbian style learning explicitly. This is especially the case given the arbitrary nature of the Hebbian learning rules used in many architectures anyway, as pointed out by Aizawa (1997a, 1997b), in the sense that those rules cannot be easily explained (or cannot be explained at all, if Aizawa is correct), or deduced, from any underlying scientific theories used contextually as part of the explanation of human cognition. Although it may seem anticlimactic to just have shown the FFN modules in S11 are in the same lot as other architectures that explicitly use Hebbian style learning, the point remains that S11 is therefore oriented towards increasing neurobiological plausibility[9].

---

[7]The interpretation of the "node and link" style of connectionist models as modelling *groups* of neurons and *groups* of synapses also raises other interesting problems that are best left as questions to be resolved in future research with connectionist architectures that implement leaky integrate-and-fire models of spiking neurons. It should be stressed, however, that *none* of the connectionist architectures cited or discussed in this thesis employ spiking neurons (except the work of Stewart & Eliasmith, 2009).

[8]It should be noted the same arguments do not necessarily apply to FFN systems that use error-backpropagation if they have many multiple layers of hidden neural units (e.g. *deep architectures*, but deep architectures like Deep Belief Nets are not necessarily trained with error-backpropagation). The above arguments apply to S11 mainly due to the fact that FFN modules in S11 have *no* hidden layers.

[9]This does not mean, of course, that S11 is any farther ahead than other purely connectionistic systems that use *only* Hebbian style learning, but given that there may be some doubts as to the neurobiological plausibility of hybrid connectionist systems that employ error-backpropagation, it is worthwhile finding that S11 is no farther behind.

5.3 Plate Focuser

5.3.1 Implementation as a Connectionist Module

The two modes of operation of the Plate Focuser described in section 4.6.3 and section 4.7.2 constitute what is implemented in the actual computer software implementation of the S11 architecture, and is one of the reasons why S11 is a *hybrid* connectionist architecture rather than a *purely* connectionistic one. I will now describe how the Plate Focuser can, in principal, be implemented connectionistically *without* having implemented a full Turing machine[10]. In the connectionistic implementation, the Plate Focuser is split into two distinct modules: one that operates during inference (call this *Plate Focuser I*), and one that operates during learning (call this *Plate Focuser L*). I will first describe the Plate Focuser I by way of a simplified case, through which it will become clear how the Plate Focuser I can be implemented.

For ease of description then, let us consider the simplified case where the scene structure has exactly two plates, and the "directions" possible are thus *self* and *other*. The Plate Indicator in this case would then only have two neural units, which we will call *binders*. Furthermore, the Plate Chooser in this case would then only have two neural units since there are only two "directions". Let us call the pair of neural activations resulting from the Plate Chooser's output-layer undergoing a winner-take-all process the *direction indicator*. During inference in this simplified case[11], the Plate Focuser I must use the two binders — one of which has an activation value of one (call this the *on state*, and we simply say that binder is *on*) and the other zero (call this the *off state*, and we simply say that binder is *off*) — and the direction indicator to reset the binders so that the appropriate binder is on

[10] Although it is possible to implement a Turing machine to do general computation using a neural network, doing so here is inappropriate as they are not generally recognized as neurobiologically plausible, given they require infinite precision and are sensitive to noise. See, e.g. Siegelmann & Margenstern, 1999; Hyötyniemi, 1996.

[11] Note that we are currently considering only a simplified case to ease the task of understanding the actual S11 implementation. In fact, the number of plates in the scene structure can be arbitrary, while the number of "directions" is, in S11, always five. The number of neural units in the Plate Indicator is thus always the number of plates in the scene structure. Further, the scene structure in S11 models just a small portion of the complete conceptual substrate that the mind operates upon, but the number of plates in the conceptual substrate is large and fixed, and thus the Plate Indicator for the whole conceptual substrate must have a fixed size. To loosely describe the conceptual substrate using the terminology of Marcus' treelets, we may say that the number of register sets that make up the mind's conceptual substrate is fixed and are arranged in a preorganized manner, while a dynamically sized treelet is instantiated amongst the register sets such that multiple treelets can be instantiated simultaneously by using non-overlapping groups of register sets.

and the other off.

Further suppose for our simplified case that for each binder, there exist a *router*, which is a sequence of neural units, with the same number of neurons as the Plate Chooser's output-layer (in this scenario, two), and with the same "directionality" interpretation as the Plate Chooser's output-layer. Suppose each router has connections from the direction indicator laid out in a *topographic* fashion: *topographic* meaning that each neural unit in the direction indicator is connected to one unit in the router, and the spatial relationship between neural units in each region (the router and the direction indicator) is preserved. Suppose each binder is fully-connected to the neural units in its associated router.

For each router, we further suppose the neural unit representing the self direction is connected to the binder associated with that router, while the unit representing the other direction is connected to the binder associated with the other plate. See figure 5.1. We will let all connections have weights of one in this simplified case. We let the activations received by the neural units in the router be *multiplied* together rather than summed, but the binders sum their received activations as usual. Let all neural units in the Plate Focuser I module have as activation function the identity function. Note that the use of multiplicative connections with the received activations is not novel within connectionist research, and has been used with success, e.g. by Hinton (Rumelhart & McClelland, 1986, pg. 115).

The Plate Focuser I would operate in the following manner during inference in our simplified case. One of the binders is initially on, and the other off. The direction indicator is set through the winner-take-all process applied on the output-layer of the Plate Chooser. Both the direction indicator and the binders would fire into the routers, and the neural units in the routers would multiply the incoming activations. Since each binder is fully connected with its associated router, all the neural units within one of the routers would receive an activation of zero from the binder that is in the off state, and therefore all the neural units in that router would have an activation value of zero (due to the multiplication of incoming activations). For the router with neural units that received an activation of one from the binder that is in the on state, it also receives activation from the direction indicator in a topographic manner, and thus, after multiplication of the incoming activations, only the neural unit that is one-to-one mapped from the direction indicator would have an activation value of one (the other units would have a value of zero). Finally, the routers fire into the binders, which sum up their received activations, and thus the binders are essentially "reset"

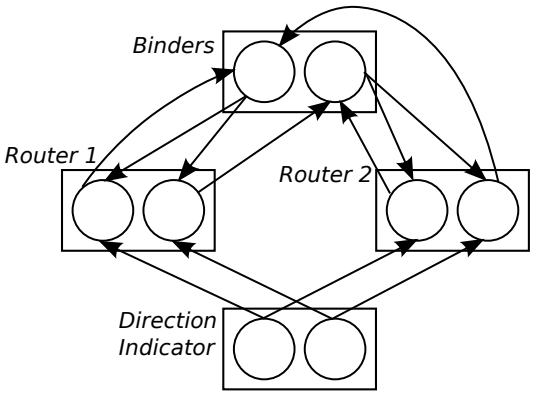

Figure 5.1: The Plate Focuser I module (for inference), consisting of routers, binders, and direction indicator, for a system with only two plates and two directions.

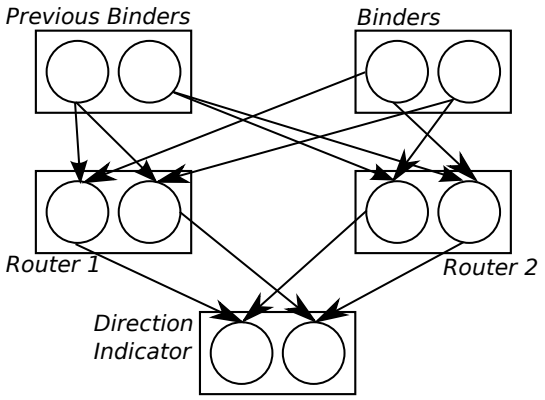

Figure 5.2: The Plate Focuser L module (for learning), consisting of routers, binders, previous binders, and direction indicator, for a system with only two plates and two directions. Connections to copy the binders to the previous binders are not shown.

so that the appropriate binder becomes on and the rest become off.

Let us now consider the Plate Focuser L module (for learning). We will continue to consider the Plate Focuser L within the above described simplified case. Furthermore, recall that we will consider the Plate Focuser L module as being completely separate and independent of the Plate Focuser I module described above. It may be possible to reduce the number of necessary neural units and to let certain connections be shared, but it only complicates the description here needlessly for the simple purpose of showing that the module can be implemented connectionistically.

The Plate Focuser L will again have a set of binders (again, two in this simplified case) and a direction indicator (again, with two neural units in this simplified case). It will also have a second set of binders (call this the "previous binders", which functions analogously to the Auxiliary Plate Indicator introduced in section 4.6.2) that serve as a copy of the first set of binders. Given the previous binders and the current binders, the Plate Focuser L must figure out which is the correct "direction" to have been taken by the Plate Chooser, i.e. figure out the correct direction indicator.

Suppose also that for each of the previous binders, we assign to it a router, which is a

sequence of neural units, with the same number of neurons as the Plate Chooser's output-layer (in this scenario, two), and with the same "directionality" interpretation as the Plate Chooser's output-layer. Each router has connections to the direction indicator in a one-to-one and topographic fashion. Each of the previous binders is fully-connected to the neural units in its assigned router (for convenience, we may also say that the previous binder is assigned to the router).

For each router, further suppose the neural unit representing the self direction is connected from the binder of the plate that is considered as being in the self direction relative to the plate associated with the previous binder which is assigned to that router. Similarly, the neural unit representing the other direction is connected from the binder of the plate that is considered as being in the other direction relative to the plate associated with the previous binder which is assigned to that router. See figure 5.2. We let all connections have weights of one. Let the activations received by the neural units in the router be *multiplied* together rather than summed, but the neural units in the direction indicator sum their received activations as usual. Let all neural units in the Plate Focuser have as activation function the identity function.

The Plate Focuser L would operate in the following manner during learning in our simplified case. One of the previous binders is initially on, and the other off. Next, suppose the Goal Object Resolver (GOR) sets one of the current binders on, and the other off. The binders and the previous binders then fire into the routers, which multiply their received activations. Since each of the previous binders is fully connected with its assigned router, and since the neural units in the routers multiply their received activations, all neural units of all routers, except the one assigned to the previous binder that is in the on state, would have a final activation value of zero. Furthermore, since only one of the binders (set by the GOR) has an activation of one, only one neural unit in the router assigned to the previous binder that is in the on state would have a final activation value of one (all others in that router would have values of zero). Specifically, the single neural unit in that router with an activation value of one is the neural unit that represents the correct "direction". The routers then fire into the direction indicator, which has neural units that sum their input activations, but since only one neural unit in any router has an activation value of one, the direction indicator will essentially just be copying the router assigned to the previous binder that is in the on state, and which contains the correct direction indicator.

Obviously, the above simplified case can be generalized to an arbitrary but fixed number

of binder nodes, and to represent five "directions" in the direction indicator. The routers would then have five neural units each, and they would simply have to be connected with the appropriate five plates that are the root, self, parent, left-child, and right-child plates relative to the plate assigned to said router. For plates that are the *leaf* plates in the binary tree structure of the scene structure (see fig 4.2), i.e. the plates that have no children, the left-child and right-child directions in the associated router could simply be connected to the self plate instead — the resulting scene structure that is built may not be technically correct, but this avoids "overflowing" the scene structure.

In any case, the scene structure here models only a portion of the conceptualization of the complete situation, i.e. only a portion of the conceptual substrate, that the mind operates upon. Thus, the memory limits of the scene structure is an artificial limitation imposed by the *architecture* here, rather than a limitation of the cognitive theory. This section constitutes a description of how the Plate Focuser module can, in principal, be implemented connectionistically *without* having implemented a full Turing machine.

5.3.2 Issues of Neurobiological Plausibility

Let us now consider whether such a connectionistic implementation of the Plate Focuser is neurobiologically plausible to any degree whatsoever. To do so, we focus only on the connectionistic implementation described in section 5.3.1, i.e. on Plate Focuser I (for inference) and Plate Focuser L (for learning).

Certainly, an argument against the Plate Focuser I and Plate Focuser L may begin by noting that the neural unit connections are seemingly ad hoc and precise, so that the connectivity seem too "hand crafted". In reply, notice that the Plate Focuser I and Plate Focuser L would seem to require no more precise an ability to wire the neural units together than what is required for connectionist systems that use holographic reduced representations (HRR) (see, e.g. Neumann, 2002). In fact, it seems that the Plate Focuser I and Plate Focuser L would require *less* precision than required for implementing HRR systems, since the Plate Focuser I and Plate Focuser L do not require the precision of wiring over possibly thousands of neural units, as is required for HRR systems. Nonetheless, the precision required for HRR systems have been criticized as being implausible biologically as well (Hadley, 2011, p. 31), so it is perhaps not good to make a comparison with HRR systems here.

Instead, lets consider in more detail what kind of seemingly ad hoc and precise connections are required. First, notice that connections between the direction indicator and the routers are topographic in nature (in both the Plate Focuser I and Plate Focuser L). It turns out that topographic connections between groups of biological neurons is pervasive in the brain (Thivierge & Marcus, 2007; Kaas, 1997), widespread for the transmission of sensory information (e.g. see McLaughlin & O'Leary, 2005; Ding & Marotte, 1997; Kunes, Wilson, & Steller, 1993; Sereno, 1991), but also extends into higher centers of neural processing (e.g. see Weisz et al., 2004; Mailly et al., 2001; O'Leary, Yates, & McLaughlin, 1999; Tamamaki & Nojyo, 1995; Tootell, Switkes, Silverman, & Hamilton, 1988), and are conjectured to be involved in "several types of complex mental operations, including reasoning and analogy making" (Thivierge & Marcus, 2007, p. 252. e.g. see Hagler & Sereno, 2006; Pulvermüller, 2005). Certainly, our widespread use of topographic connections, between groups of neurons in both the Plate Focuser I and Plate Focuser L, should not worry us too much in consideration of their neurobiological prevalence.

Now, consider the one-to-many connections from the binders to each router in the Plate Focuser I, and from the previous binders to each router in the Plate Focuser I. It turns out that one-to-many connections (i.e. *divergent* connections) are also frequently empirically found throughout the brain (Thivierge & Marcus, 2007, p. 251. e.g. see Morel, Liu, Wannier, Jeanmonod, & Rouiller, 2005; Rockland & Drash, 1996; Saleem & Tanaka, 1996). Furthermore, they are used pervasively in various connectionist networks, such as Deep Belief Nets (DBN), and there is evidence that certain DBN architectures work in a manner similar to the standard model of the visual cortex (Hinton, 2010, p. 183). Thus, from the standpoint of connectionism, which is committed to building models that are neurobiologically inspired, even if not neurobiologically realistic, our widespread use of divergent connections between groups of neurons in the Plate Focuser I and Plate Focuser L should not worry us too much either.

Finally, consider the connections between the binders and the routers that are *not* divergent connections (in both the Plate Focuser I and Plate Focuser L). I must admit that they do seem ad hoc in nature, but consider that there are only *five* such connections per router that requires justifying, as opposed to, e.g. the possibly thousands required for HRR systems, the problem is smaller than it first appears. Furthermore, recall from section 4.6.3 that the *functionality* implemented by the Plate Focuser (together with the Plate Chooser) at a higher level of description is an instance of the sequential scanning

cognitive phenomenon that is employed, e.g. for "mentally tracking an event as it unfolds through time, that is, scanning sequentially through it along the temporal axis" (Langacker, 2008, p. 111); and is also employed in apprehending the nested locative construction, e.g. as appears in the sentence, "Your camera is upstairs, in the bedroom, in the closet, on the shelf" (ibid., p. 195), where one must mentally locate the trajector (the camera) by "successively 'zooming in' to smaller and smaller areas" (ibid.). Thus, the existence of the Plate Focuser, even if it is not implemented in exactly the manner proposed here, is supported in a "top-down" manner, by appealing to theories of conceptual semantics and Cognitive Grammar.

Therefore, although the aforementioned five connections per router appear ad hoc, their connectivity is reasonable within the context of supporting the functioning of a cognitive module that is supported by cognitive theory. Before leaving this topic of justifying those five connections per router (that do not fit the aforementioned divergent or topographic pattern of connectivity), let us reconsider their function. In the case of the Plate Focuser L, the five connections are from the binders to the routers, and neural units in the routers can be seen as performing a "gating" function: i.e. a neural unit will pass on neural activation from the previous binders to the direction indicator only if the binder connected to it is on.

It turns out this is an instance of what is known as *permissive synaptic gating* in the brain. Biological neurons that do this kind of gating (e.g. *bistable neurons* and other so-called *"up/down" neurons*) have been observed to operate in various regions of the brain, including regions associated with language processing, and there is evidence that they are "capable of controlling the flow of information from one set of neurons to another" (Gisiger & Boukadoum, 2011, p. 1. See also P. S. Katz, 2003; O'Donnell & Grace, 1995). Of course, within the Plate Focuser I, the five connections per router, from the routers to the binders, may also be interpreted similarly as participating in a synaptic gating function (e.g. by introducing a separate cyclically firing node that fires into the binders, and by making the binders into *"up/down" neurons* whose "up/down" state is set by the activations from the routers, so that only one binder is ever in the "up" state. When the separate cyclically firing node fires into the binders, only the binder in the "up" state would be able to pass on the neural activation to the router; the binders in the "down" state do not fire).

It now seems the Plate Focuser I and Plate Focuser L may be more neurobiologically plausible than it appeared from our initial assessment of the situation. All in all, especially

given that a great deal of connectionist research is concerned with producing neurobiologically *inspired* models of cognition, and not necessarily neurobiologically realistic models, it seems evident that the Plate Focuser module can be implemented connectionistically such that it is oriented towards increasing neurobiological plausibility.

5.4 Information Channels

Let us now discuss the neurobiological plausibility of the Information Channels module of our connectionist system, which is a term introduced to refer to all the copying of activation values that is necessary between all the other modules in S11. For instance, the hidden-rep inferred by the Lexicon Module is copied into the input-layer of all the other three feedforward network modules. As discussed above, the Plate Chooser also requires the entity-rep stored in every cell in the plate in focus in the scene structure to be copied into its input-layer. All this copying is performed in the software implementation of our connectionist model by ordinary programmatic methods (e.g. copying the floating point numbers representing neural activations from one array to another), and in part because of this, the implemented model is a hybrid connectionist model.

Consider how the Information Channels may be implemented in a purely connectionistic fashion. That should be fairly simple now, in view of the biological neural network discussion above. We simply need topographic connections between all the sets of neurons that may be copied to or from, and then place "up/down" neurons between all of the previous topographic connections so that they may be used to control which specific set of neural activations gets copied to which other set of neural units. In fact, just such a basic gating mechanism, amongst several others, has been investigated recently, and there is preliminary evidence that they exist in areas of the brain, including the prefrontal cortex (Gisiger & Boukadoum, 2011, pps. 3, 7), which has been indicated as an area that participates in semantic processing (Fiez, 1997).

An objection may be raised regarding the extent of connections required, and whether it is reasonable to fashion so many connections for transferring information between the different modules. Consider, however, that "[w]hite matter makes up about 50% of human brain volume" (Filley, 2010, p. 159), and "subcortical white matter (WM), which contains the axons that interconnect nearby as well as distant areas in the [cortical gray matter] and their subcortical targets... comprises... >40% of the cerebral cortex of dolphins,

whales, elephants, and humans" (Herculano-Houzel, Mota, Wong, & Kaas, 2010, p. 19008). Furthermore, "[i]t is estimated that the great majority of white matter fibers connect different cortical regions rather than connect the cortex and subcortical structures" (Zhang & Sejnowski, 2000, p. 5621). Since "a general conceptual formulation is that white matter supports information transfer to complement the information processing carried out by gray matter" (Filley, 2010, p. 158), it would not be too preposterous a proposition that the Information Channels are implemented as white matter, and so the amount of connections required for the Information Channels between modules is actually not implausible.

Another objection may still be raised that even though the small-scale "hand crafted" connections may be explained as pervasively occurring patterns of neural connectivity in the brain, the large-scale structure of the Information Channels connecting the various modules is not neurobiologically plausible, as they too seem overly "hand crafted". Amazingly, there is in fact evidence from fMRI studies on infants that "the infant brain does not respond to speech in diffuse areas, as connectionist models, which suggest in their extreme form an equipotential brain, would have predicted... Rather, it is functionally structured, recruiting distant regions in cooperative networks... [T]he processing properties of the infant brain... make it efficiently adapted to the most frequent auditory input encountered by the human infant, namely speech... These capacities rely mostly on brain circuits similar to those observed in adults" (Dehaene-Lambertz, Hertz-Pannier, & Dubois, 2006, p. 372). It may well be that the large-scale structure of the brain is pre-determined genetically and is not learned from data (except on an evolutionary time scale). Evidence for this contention may be found, for example, in the case of an experiment with mice where a certain protein is deleted so that neurotransmitters required for the maintenance of synaptic connectivity and synaptic transmission is not produced in the brain, and yet, even in the absence of synaptic transmission, that did not "prevent normal brain assembly, including formation of layered structures, fiber pathways, and morphologically defined synapses" (Verhage et al., 2000, p. 864). That is to say, in the mice without neural transmission, the area of the brain that usually develops at a given embryonic stage of development[12] was indistinguishable from the same area of the brain of a normal mice, and this was true also at birth (ibid.,

[12]The comparison is restricted to the area that usually develops at a given stage of development because neurons in previously developed areas die off without the required neurotransmitters, thus showing that the neurotransmitters required for synaptic transmission is required for the *maintenance*, but not the initial development, of the brain.

p. 866).

All in all, especially given that a great deal of connectionist research is concerned with producing neurobiologically *inspired* models of cognition, and not necessarily neurobiologically realistic models, it seems evident that the Information Channels module is oriented towards increasing neurobiological plausibility.

Chapter 6

Experiments and Results

To provide an interesting comparison of the abilities to exhibit systematicity and assign meaning to sentences, the experimental results will be collected using a similar testing regime as employed by Hadley and Hayward (1997) and Hadley and Cardei (1999); furthermore, there will be two sets of sentences that system S11 will be trained and tested on (i.e. two different corpora), where one corpus will be generated by a simple recursive grammar similar to that used by Hadley and Hayward (1997), and the other will be generated by a simple recursive grammar that is an augmentation of the one used by Hadley et al. (2001).

It may seem strange that this thesis will show systematicity properties in S11 through experimentation on artificial languages produced by simple recursive grammars. It is, however, important to stress that quite a few connectionist implementations are of this kind within cognitive science, as demonstrated by all the works surveyed in sections 2.3 and 2.7, and also the works of, e.g. Neumann (2002); Elman (1991); and Elman (1990).

The corpus will be separated into a training and a testing set. The training set will have certain classes of sentences excluded in order to check for the generalization abilities of S11. The specifics of each of the two corpora, including results of experiments performed on them, will be specified in sections 6.1 and 6.2. The S11 system is allowed to learn the training set of sentences, and once learning is complete, it will be tested on the testing set of sentences.

The testing procedure itself involves presenting each of the test sentences to S11, word by word. As each word is presented, it is converted to its corresponding word-rep (which, as you will recall from section 4.5, wherein the manner in which word-reps are produced

for a given word is specified, models the required phonological cognitive structure), and the word-rep is processed by S11 using the inference procedure described in chapter 4. As each word-rep is processed, the system will progressively create a corresponding scene structure as per the inference procedure described in section 4.6. When the entire sentence has been processed, the scene structure is saved for analysis.

Analysis occurs by comparing the inferred scene structure that was saved against the correct scene structure for the corresponding sentence. The correct scene structure for a sentence is constructed in the same way it was constructed for the purpose of training (see section 4.3). The comparison will be made by *flattening* the inferred and correct scene structures into vectors, and then calculating the *cosine similarity*, i.e. the cosine of the angle, between the two resultant vectors[1]. Cosine similarity is a good measure of how close two vectors are, where a cosine similarity score of one means the vectors are identical, and a score of zero means the vectors are orthogonal.

Notice that the cosine similarity as defined would be undefined if one or both of the vectors is a zero vector. To smooth out this unfortunate and rare wrinkle, just in case the system infers a scene structure that flattens out into a zero vector, we make the following two stipulations in accordance with intuition: First, we stipulate that if only one of the two vectors compared is a zero vector, then the cosine similarity between them is defined to be zero. Second, we stipulate that if both vectors being compared are zero vectors, then the cosine similarity between them is defined to be one. Thus nothing is similar to the zero vector other than itself.

Since the plates of a scene structure have a binary tree structure, each plate contains multiple cells, and each cell contains a sequence of neural units, the process of flattening a scene structure into a vector needs to be defined precisely. We will define *flattening* a scene structure into a vector as follows:

1. Let E be the entity-rep from cell C_n, then $V(C_n)$ is the vector representation of E, and is simply the sequence of neural activations, encoded as a sequence of floating point numbers, stored as E.

2. Let $C_{1,m}, C_{2,m}$, and $C_{3,m}$ be the three cells from plate P_m, then $V(P_m)$ is the vector formed by concatenating the vectors $V(C_{1,m}), V(C_{2,m})$, and $V(C_{3,m})$, in that order

[1]For convenience, we may speak of the cosine similarity of two scene structures when in fact, the cosine similarity is a measure of the two vectors resulting from flattening the two scene structures.

(this means if $V(C_{i,m}) \in \mathbb{R}^N$, then $V(P_m) \in \mathbb{R}^{3N}$).

3. Let $(P_1, \ldots, P_M)$ be a sequence of the plates from the scene structure, taken in a breadth-first traversal pattern of the binary tree structure of the plates, then the concatenation of vectors $V(P_1), \ldots, V(P_M)$, in that order, is defined as the resultant vector from flattening the scene structure.

The learning procedure is as described in section 4.7, but a question arises as to whether we allow the system to learn every single sentence in the training set, and how many epochs of training do we allow the system to use. What we do here is a kind of *early-stop training*, where the training set is further randomly partitioned into a *training subset* and a *validation subset*. The system is allowed to learn the sentences of the training *subset* for as many epochs as necessary to achieve an acceptable performance level as measured by the average cosine similarity of the scene structures produced from processing sentences in the validation subset. In practice, we allow fractional epochs, i.e. the system will train on blocks of 25 sentences from the training *subset*, and will be tested on the validation subset after training on each block; learning is complete when the cosine similarity score on the validation subset is above a threshold or has converged. The threshold will be set to 0.97. Further, the size of the *validation subset* is set to be 20% of the training set; the remaining 80% of the training set will be the *training subset* (Note the difference between the training *subset* in comparison to the complete training *set*. It would be incorrect to say that the validation subset is not part of the actual training set as S11 uses the validation subset during training, although it does not use it directly in error-backpropagation).

The experiments were conducted on a MacBook Pro using 4 GB of RAM and a 2.8 GHz Intel Core i7 processor. The software was written in Clojure (version 1.2.0), a dialect of the Lisp programming language that runs on the Java Virtual Machine (version 1.6.0).

6.1 Corpus 1

Corpus 1[2] is generated by a syntax (call it *Syntax 1*) that is an augmentation of the one used by Hadley et al. (2001). The testing of the system using Corpus 1 is intended to show the straightforward generalization ability of the system in the tradition of standard

[2]The specific set of sentences that constitute Corpus 1 is available from the author upon request.

machine learning literature. That is to say, Corpus 1 is a set of randomly generated sentences generated by Syntax 1, and it is randomly partitioned into a training and testing set without any attempt at specifying special restrictions as to what class of sentences are disallowed from the training set. Therefore, we are not explicitly testing the ability of the system to exhibit systematicity, but simply testing its ordinary generalization ability. If the system performs poorly on Corpus 1, then there would be no hope of the system exhibiting any level of systematicity at all.

Corpus 1 contains 1300 unique sentences. It is randomly partitioned into a training set containing 1000 sentences, and a testing set containing 300 sentences. Note that due to the random partitioning of the training set into training and validation subsets, the system is in fact only learning 800 sentences by error-backpropagation, and using the remaining 200 sentences to validate that learning has occurred sufficiently.

Syntax 1 (see figure 6.1), besides changing some of the words used (a trivial change), extends the syntax used by Hadley et al. (2001, p. 76) by allowing each noun to optionally be preceded by an adjective, and by allowing each verb to optionally be preceded by an adverb. Syntax 1 is an unambiguous grammar, i.e. each sentence generated has exactly one derivation. Note that Hadley et al. (2001) created a connectionist system that predicted the semantics of the *next word*, but does not try to "build up" a semantic representation of the scene described by the whole *sentence* (recall the discussion from section 2.8). Therefore, it is not possible to compare the performance of the architecture created by Hadley et al. (2001) with the performance of the S11 architecture created in this thesis, since they are designed simply to do different things.

After learning is completed, the S11 system was able to achieve, as averaged over five experiments, a performance of 0.973 cosine similarity averaged across all test set sentences, and learning required an average of 8.78 epochs (which took, on average, approximately 83 minutes to complete). This result is robust as similar performance was observed over more than five experiments, and the system never failed to achieve a validation cosine similarity score above the threshold.

As a point of comparison, after learning is completed, the S11 system was able to achieve, as averaged over five experiments, a performance of 0.975 cosine similarity averaged across all *training* set sentences. Also, *prior to any learning taking place* (i.e. while the connection weights in the FFN sub-modules of S11 were still in their randomly set state), the S11 system was able to achieve, as averaged over five experiments, a performance of only 0.147

S $\longrightarrow$ 'start' NP VP NP

NP $\longrightarrow$ N | N RC | N PP | ADJ N | ADJ N RC | ADJ N PP

N $\longrightarrow$ 'women' | 'girls' | 'birds' | 'bb-bats' | 'men' | 'boys'
| 'chairs' | 'baseballs' | 'dogs' | 'tables' | 'cats' | 'mice'

ADJ $\longrightarrow$ 'red' | 'green' | 'blue' | 'heavy' | 'light'
| 'tall' | 'short' | 'furry' | 'smooth' | 'hairy'

VP $\longrightarrow$ V | ADV V

V $\longrightarrow$ 'chase' | 'sees' | 'swing' | 'love' | 'avoid'
| 'follow' | 'bump' | 'hit' | 'consume' | 'dislike' | 'like'

ADV $\longrightarrow$ 'physically' | 'rapidly' | 'slowly' | 'smoothly' | 'jumpily' | 'nicely' | 'badly'

RC $\longrightarrow$ 'that' VP NP | 'that' N VP | 'that' ADJ N VP

PP $\longrightarrow$ 'from' NP | 'with' NP

Figure 6.1: Syntax 1. For convenience, we use 'start' to denote the beginning of a sentence, rather than use '.' to denote its end.

S ⟶ 'start' NP V NP
NP ⟶ N | N RC
N ⟶ 'women' | 'girls' | 'birds' | 'bb-bats' | 'men' | 'boys'
| 'chairs' | 'baseballs' | 'dogs' | 'tables' | 'cats' | 'mice'
V ⟶ 'chase' | 'sees' | 'swing' | 'love' | 'avoid' | 'follow' | 'bump' | 'hit'
RC ⟶ 'that' V NP

Figure 6.2: Syntax 2. For convenience, we use 'start' to denote the beginning of a sentence, rather than use '.' to denote its end.

cosine similarity averaged across all of the sentences in the *validation* subset. This shows learning indeed occurred.

6.2 Corpus 2

Corpus 2[3] is generated by a syntax (call it *Syntax 2*, and see figure 6.2) that is meant to resemble closely the one used by Hadley and Hayward (1997). Syntax 2 is an unambiguous grammar, i.e. each sentence generated has exactly one derivation. The testing of the system using Corpus 2 is intended to show the systematicity capacity of the system. As such, Corpus 2 consists of a testing and training set of sentences, where the training set is generated with certain special classes of sentences excluded (i.e. it is relevantly sparse). Therefore, we are explicitly testing the ability of the system to exhibit certain forms of linguistic systematicity, a stronger form of generalization ability than what was tested with Corpus 1 (recall the discussion on systematicity from section 2.4).

Specifically, the training set contains 685 unique sentences. Of the 685 sentences, 512 sentences (i.e. 74.7% of training set) are simple noun-verb-noun sentences, and 173 sentences (i.e. 25.3% of training set) are complex sentences containing relative clauses with an embedding depth of two. Of the 173 complex sentences, half of them (i.e. 86 sentences) contains a single relative clause with a single nested relative clause, while the other half (i.e. 87 sentences) contains, for both subject and object nouns, a relative clause with a single

[3]The specific set of sentences that constitute Corpus 2 is available from the author upon request.

nested relative clause (i.e. four relative clauses in total, but only to embedding depth of two).

Furthermore, all twelve nouns appear in the training set, but only four of them are permitted to appear in both subject and object positions. Of the remaining eight nouns, four appear only in the subject position, and four appear only in the object position. Note that due to our random partitioning of the training set into a training and a validation subset, the system is in fact only learning 548 sentences by error-backpropagation, and is using the remaining 137 sentences to validate that learning has occurred sufficiently.

The test set for Corpus 2 contains 5134 randomly generated unique sentences. Of the 5134 sentences, 640 are simple sentences of the noun-verb-noun variety. Another 1494 of the 5134 sentences present some noun in a position it did not occupy during training; of the 1494 sentences, two-thirds contain embedded clauses up to a depth of three in either the subject or the object noun-phrase, and the remaining one-third contain embedded clauses up to three levels deep in both the subject and the object noun-phrases. The remaining 3000 sentences of the 5134 are randomly generated using Syntax 2 with no restrictions placed on any word's syntactic position, with embedded clauses up to a depth of three.

With these restrictions in place during the generation of the training and testing sets of sentences, Corpus 2 is thus relevantly sparse with respect to semantic systematicity. Therefore, if S11 performs successfully on the testing set after learning the training set, the system can then be said to meet the requirements for semantic systematicity (Hadley & Hayward, 1997). Note that without these, or similar, restrictions in place when generating the training set, the ability for a system to exhibit semantic systematicity would simply not be tested at all — i.e. these restrictions are designed specifically to test a system's ability to exhibit semantic systematicity, as required by the definition of "semantic systematicity" (c.f. section 2.4).

In comparison to the training and testing corpora used by Hadley and Hayward (1997), the significant difference of Corpus 2 is that the training set here allows embedding up to level two, whereas Hadley and Hayward allowed embedding only up to level one. Further, Hadley and Hayward also generated 872 more sentences for their test set.

In terms of performance, the connectionist system created by Hadley and Hayward (1997) is capable of correctly processing every single one of the sentences in the test set they used. It is an impressive ability, indeed, but we should keep in mind the problems and limitations of the system as discussed in section 2.8, and especially the problems raised

by Aizawa (1997a, 1997b) and the variant criticism as discussed in section 2.8.1. It is also important to emphasize[4] that the actual corpus of sentences used for training and testing by Hadley and Hayward (1997) is *different* than Corpus 2, and that the meaning representation used by Hadley and Hayward is *drastically different* than the scene structure used in S11; therefore, it can be argued that the performance results of the architecture by Hadley and Hayward, and of the S11 architecture, are *incommensurable*.

After learning is completed, the S11 system in this thesis was able to achieve, as averaged over five experiments, a performance of 0.970 cosine similarity averaged across all test set sentences, and learning required an average of 1.99 epochs (which took, on average, approximately 37 minutes to complete). This result is robust as similar performance was observed over more than five experiments, and the system never failed to achieve a validation cosine similarity score above the threshold.

As a point of comparison, after learning is completed, the S11 system was able to achieve, as averaged over five experiments, a performance of 0.981 cosine similarity averaged across all *training* set sentences. Also, *prior to any learning taking place* (i.e. while the connection weights in the FFN sub-modules of S11 were still in their randomly set state), the S11 system was able to achieve, as averaged over five experiments, a performance of only 0.146 cosine similarity averaged across all of the sentences in the *validation* subset. This shows learning indeed occurred.

To further the analysis of the performance of S11, consider that more detailed analysis of the experimental results show that, *averaged over five experiments, on 5.9% of the test set sentences the S11 system achieved a cosine similarity score of* one *(i.e. a perfect score).* As for the other 94.1% of the test set sentences (i.e. the sentences on which S11 did not achieve a one cosine similarity score), averaged over five experiments, the S11 system achieved a cosine similarity score of 0.968.

What is more, the Plate Chooser and Cell Chooser modules never *made an error during testing.* That is to say, for each word in *every* test sentence, the Plate Chooser allowed the S11 system to *always* figure out correctly which plate to focus on in the scene structure, and the Cell Chooser allowed the S11 system to *always* figure out correctly which cell to focus on in the plate in focus. So the non-one cosine similarity scores from the aforementioned 94.1% of test set sentences was due entirely to the Entity Compositor's fault.

[4]Thanks to A. Sarkar (personal communication, 27 June 2011) for pointing out this issue.

Since, for each word processed, S11 always figured out correctly which plate and cell to focus on, we can compare (e.g. by cosine similarity) how close the Entity Compositor's inferred entity-rep is with the correct entity-rep stored in the cell in focus, in the plate in focus, in the clamped scene structure. In fact, averaged over five experiments, the Entity Compositor achieved a cosine similarity score of 0.956 averaged across all words in all test set sentences, when comparing the entity-rep inferred by the Entity Compositor with the correct entity-rep.

It may be argued[5] that since the scene structure is bound to contain many unused cells when processing a sentence, the vector resulting from flattening the scene structure may contain many zeros, and these zeros may artificially and unfairly inflate the cosine similarity score. Fortunately, this is not the case in the experiments, as shown by the fact that the Entity Compositor achieved a high cosine similarity score when examined on its own, since by directly comparing entity-reps, the unused cells do not factor into the cosine similarity score for the Entity Compositor at all. Furthermore, it should be reemphasized that the Cell Chooser and Plate Chooser modules never made an error during testing. Notice also that the error introduced by the Entity Compositor has, in the S11 system, little to nothing to do with systematicity and the processing of embedded clausal structures.

Recall that the construction of the Corpus 2 training and testing sets specifically tests the ability of S11 to display strong and semantic systematicity. The experimental results show that S11 is capable of processing successfully a large fraction of the testing set. Furthermore, the zero error displayed by the Plate Chooser and Cell Chooser modules show that S11 is capable of processing perfectly the clausal structure of Syntax 2, despite the Entity Compositor's problem with inferring the entity-rep associated with a word-rep. Taking into account the fact that S11 was tested on 5134 sentences after training directly on only 548 sentences that were generated with a highly restricted version of Syntax 2, it is fair to say that, with respect to Corpus 2, S11 exhibits strong systematicity to a significant degree, and also exhibits partial semantic systematicity (and recall that these are terms of art previously defined in section 2.4, especially noting that these terms of art do not cover all possible levels of linguistic systematicity or semantic ambiguity, as discussed also in that section).

[5]Thanks to A. Sarkar (personal communication, 3 August 2011) for pointing out this issue.

6.3 Bootstrapped Learning: Preliminary Results

Although the issue of bootstrapping or interactive learning is a complex issue that will require further research in the future, some preliminary experiments on a modification of S11 has already shown the promise of a certain form of bootstrapped learning. So much so that the following preliminary results are worth sharing as part of this thesis.

Recall that in S11, during the entity focusing process of the learning phase, the GOR enables S11 to identify and focus on the correct entity-rep modelling the conceptualization of the goal object. Since the GOR is assumed to be able to identify the goal object with complete accuracy, the learning phase's entity focusing process is also assumed to be infallible, because the entity focusing process in S11 relies completely on the GOR. Such an entity focusing process during learning is surely an idealization only.

It turns out, however, that a bootstrapping process can in fact be employed during the learning phase's entity focusing process, so that the GOR need not be used all the time during learning, thereby modelling a more cognitively realistic process of gradual bootstrapped learning of the entity focusing process. The changes to S11 necessary to implement the bootstrapping of the learning phase's entity focusing process results in a modified architecture, which will be called *S11b*.

The following will specify the two modifications necessary to S11 for the creation of S11b, and the preliminary experimental results of S11b on Corpus 2 will be presented afterwards. The inference process of S11b will be exactly the same as in S11, but the learning process is modified in the following two ways:

1. Recall from section 4.7.3 that the Plate Indicator and Cell Indicator are set during the learning phase's forward-pass, and is effectively a guess that a cell g is the cell in focus and that the plate containing g is the plate in focus. Subsequently, as part of the entity focusing process of the backward-pass in S11, the GOR indicates that a cell j contains the correct entity-rep to focus on by setting the Cell Indicator and Plate Indicator appropriately.

   During the entity focusing process of the backward-pass in S11b, however, only with probability p will the GOR be used to indicate that the cell j contains the correct entity-rep to focus on; as before, the GOR does this by setting the Cell Indicator and Plate Indicator appropriately. When the GOR is *not* used, S11b will instead leave the Plate Indicator and Cell Indicator *unchanged* from what the forward-pass had set

it to; this means letting cell g remain in focus (obviously, this occurs with probability $1 - p$, given the GOR is used with probability p).

2. Notice that when cell g is indicated, during the learning phase's forward-pass in S11b, as being the cell in focus in the plate in focus, it is in fact possibly wrong (i.e. $g \neq j$, where j is the cell containing the correct entity-rep to focus on). Thus, the entity focusing process is fallible in S11b, since the GOR is used to indicate cell j is in focus only with probability p, and cell g is indicated as being in focus with probability $1-p$. Furthermore, the cell g may, in fact, be in a plate that is not even a neighbour of the previous plate in focus (as indicated by the Auxiliary Plate Indicator), whereas the Plate Chooser and Plate Focuser, operating together during inference, can *only* ever indicate a plate as the next plate in focus if that plate is a neighbour of the current plate in focus (recall section 4.6.3). Thus, it is possible that S11b will guess a cell g to be in focus, during the learning phase's entity focusing process, that is not even in a valid plate in relation to the previous plate in focus.

   Therefore, the Plate Focuser in S11b will, during learning, also be programmed to detect whether a cell indicated as being in focus by the entity focusing process is in fact in a plate that is a neighbour of the previous plate in focus. If the indicated cell g is in a neighbouring plate (whether or not g contains the correct entity-rep to focus on is irrelevant, and thus g may be the wrong cell to focus on), S11b will continue with the backward-pass of learning as usual. If, however, g is *not* in a neighbouring plate (even if g *does* contain the correct entity-rep to focus on), the learning process is aborted for the current word being learned as well as for all subsequent words in the sentence currently being learned by S11b. Learning continues with the next training sentence instead.

The above two modifications to S11 results in the S11b architecture, which has a fallible entity focusing process as part of the learning phase (whereas S11 has an *in*fallible entity focusing process).

The S11b architecture was experimented on with Corpus 2, with the parameter p above set to various values, and the resulting actual percentage of words p_{actual} on which the GOR was used, during the learning phase, tracked. Table 6.1 contains the values of p used, the values of p_{actual} observed, the cosine similarity performances c_{test} of S11b as averaged across all test set sentences after learning is completed, and the cosine similarity

p	p_{actual} (%)	c_{train}	c_{test}	Epochs
0.3	31.0	0.164	0.129	0.51
0.4	36.7	0.991	0.976	4.47
0.5	47.1	0.983	0.973	2.87
0.6	57.5	0.977	0.966	2.51
0.7	69.6	0.977	0.966	2.28
0.8	79.7	0.978	0.969	2.10

Table 6.1: Performance of S11b on Corpus 2. p is the probability the GOR is used in the process of learning a word. p_{actual} is the percentage of words S11b actually learned with the GOR used. c_{train} is the cosine similarity performance of S11b averaged over all training set sentences. c_{test} is the cosine similarity performance of S11b averaged over all test set sentences. "Epochs" is the number of epochs S11b required during learning to reach the early-stop training's stopping condition.

performances c_{train} of S11b as averaged across all training set sentences after learning is completed.

The results in table 6.1 show that S11b was able to learn successfully even when the GOR was used during learning on only 36.7% of the words learned, but the number of epochs required is over two times that of when the GOR was used on 79.7% of the words learned. When the GOR was used during learning on only 31.0% of the words learned, the early-stop training's stopping condition, namely that of convergent cosine similarity performance on the validation subset of sentences, was reached *much* too early and the system failed to learn as demonstrated by its low c_{train} value. Although these results must be treated as preliminary, given that the architecture of S11b was not rigorously examined in this thesis in terms of neurobiological or psychological plausibility, etc., it is still encouraging to see that some form of bootstrapped learning is indeed possible within the general conception of the S11 framework.

Chapter 7

Conclusion

7.1 Discussion

The research problem stated in chapter 3 has been solved with the architecture developed in chapter 4. Consider the following:

As experimentally shown in chapter 6, the S11 architecture achieves strong systematicity, and also achieves partial semantic systematicity; therefore, the S11 architecture is behaviourally adequate with respect to Syntax 1 and 2 in modelling the assignment of meaning to input sentences (thus satisfying the research question's points 3 and 4 from chapter 3).

The S11 architecture implements a model of linguistic meaning that is rooted in Quine's concept of affirmative stimulus meaning, as shown in section 4.1, and further structured as conceptualizations as analyzed in the conceptual semantics of Cognitive Grammar, as shown in sections 4.2 and 4.3. As a model, therefore, the S11 architecture can be said to be oriented towards increasing psychological plausibility, and in view of the analysis from sections 2.5 and 2.8, it is certainly more psychologically plausible than the connectionist architectures surveyed in sections 2.3 and 2.7 (thus more than satisfying the research question's point 2 from chapter 3).

The S11 architecture is a hybrid connectionist system, but as shown in chapter 5, the modules that make it a hybrid, rather than a non-hybrid, connectionist network have been shown to be *either* capable of being implemented using purely connectionistic processing, possibly found in the brain in terms of evidence from neuroscience, *or* plausibly found as a

basic or primitive functionality of the mind in terms of evidence from psychology. Significantly, S11 does not employ any classical message passing systems of computation (notably in contrast to the architecture by Hadley & Cardei, 1999), and the error-backpropagation employed can be converted to a Hebbian style learning algorithm in future work[1], as already mentioned in section 5.2. Therefore, the overall model developed in this thesis is oriented towards increasing neurobiological plausibility (thus satisfying the research question's point 1 from chapter 3).

Since the S11 architecture is experimentally shown to be behaviourally adequate to the degree described above on the basis of the hypothesized model of linguistic meaning, and connectionistic or neurobiological processing (as described above), we therefore have confirmatory evidence that corroborates, to that degree, the hypothesized S11 architecture and the conceptual semantic theory on which it is based (as promised in chapter 3). This means, in summary, that to the extent of support from the experiments described above, the evidence suggests understanding (1) that language meaning is in part holistic; (2) that meaning is in part also conceptual, systematic, and structured in ways similar to how it is analyzed in Cognitive Grammar; (3) that language comprehension in part need not require explicit syntactic mental structures of the variety advanced by generative linguists (a claim advanced by Cognitive Grammar as well); and (4) that the ability to identify common goal objects may very well play a large role in language learning. Furthermore, S11 therefore demonstrates one possible way of providing a processing model of conceptual semantics, largely as proposed in Cognitive Grammar, that can be implemented within a connectionist framework (as promised in chapter 3).

Finally, the S11 architecture provides a possible *explanation* (and not just an exhibition) of the phenomenon of linguistic semantic systematicity in humans: language processing is semantically systematic because human *conceptualization* of the external world is itself systematic, compositional[2], and any fragment of conceptualization can be identified and manipulated independently and in isolation from all other fragments.

It is easy to see why this is a bona fide scientific explanation. Firstly, consider that

[1]Whether any specific Hebbian style learning rule is neurobiologically realistic is a question for future research. It should be noted that as analyzed by Aizawa (1997a, 1997b), at least some Hebbian style learning rules might simply be too arbitrary to be considered justifiable or explainable from the underlying scientific theories used contextually as part of the explanation of human cognition.

[2]Recall the discussion on conceptual systematicity and compositionality from section 4.3

the Plate Chooser, Cell Chooser, and Entity Compositor all function[3] on the basis that the correct "location" for an entity-rep within the space of meaning fragments can be "picked out"[4] from within the scene structure so that the entity-rep in that location can be manipulated and transformed, independently and in isolation from all other entity-reps in the scene structure, by the Entity Compositor in accordance with the word currently being processed by the system (let us call this capability *isolated transformability*). In the S11 architecture, isolated transformability is implemented by exploiting the scene structure's conceptual systematicity and compositionality.

Secondly, the experimental results clearly demonstrate the linguistic systematicity capabilities (with respect to Syntax 1 and 2) of the S11 architecture as a result of the learning regime employed. But since the functioning of the S11 architecture depends so strongly on conceptualizations being systematic, compositional, and isolated transformable, these three properties probably play a large and crucial role in making linguistic systematicity possible. We thus have a bona fide scientific explanation[5] of the semantic systematicity of language (at least with respect to Syntax 1 and 2).

In contrast to the architecture by Hadley and Hayward (1997)[6], by leaning on results from other cognitive or neuroscientific theories (and then letting *those* theories be explained by others through some evolutionary account[7]), the possible explanation of linguistic systematicity in humans offered above avoids the variant form, from section 2.8.1, of the criticism initially advanced by Aizawa (1997a, 1997b) that the explanatory power of connectionism is too weak and not very satisfying due to the overwhelming number of ad hoc or auxiliary hypotheses required, and due to the overwhelming explanatory distance between the architecture and the underlying scientific theories invoked contextually as part of the explanation.

---

[3]Recall the metaphorical intuitive descriptions of the functioning of those modules from sections 4.6.3, 4.6.5, and 4.6.6.

[4]Note that in this context of "picked out", the modules of S11 being referred to are the Plate Chooser, Cell Chooser, and Entity Compositor, and *not* the Goal Object Resolver.

[5]Recall that the emphasis on explanatory power is in view of the discussion from section 2.8.1.

[6]It is interesting to note that Hadley and Hayward (1997) only claimed to offer an explanation of systematicity "in some *possible* cognitive agent" (R. F. Hadley, personal communication, 24 June 2011) that might not necessarily be human (ibid.), and that might be existing in a possibly *different* world (R. F. Hadley, personal communication, 30 June 2011). This thesis, however, is interested in furthering an understanding of a *specific* species of cognitive agents here on Earth, namely, humans.

[7]Recall that this idea of *sharing* the responsibility of explaining phenomena is not revolutionary and is well defended. See footnote 25 from section 2.8.1 for details.

For example, note that Hadley and Hayward (1997) used a structure, for representing meaning of sentences, that assumes meaning is propositional (ibid., p. 6), that assumes the learner possesses ahead of time a "combinatorially adequate conceptual scheme" (ibid., p. 33), which might be interpreted to mean that it is assumed to be conceptually systematic, and that is assumed to "innately" (ibid.) accommodate thematic categorizations of at least agent, patient, and action (ibid.). They, however, admit that these are mere suppositions (ibid.), and no evidence is presented that meaning representations are propositional and structured in the way that they had presented it. In contrast, the scene structure developed in sections 4.2 and 4.3 model imagistic and experiential conceptual cognitive structure, does not require an innate ability to categorize objects as agent, patient, or action, and as already mentioned, finds explanatory and evidentiary support from conceptual semantic theories of Cognitive Grammar and cognitive linguistics.

Of course, the S11 architecture developed in this thesis, like all scientific theories, also requires many auxiliary hypotheses, but as an explanation for linguistic systematicity in humans, and based on the evidence provided above, the S11 architecture should at least be seen as a *better* scientific explanation than the one offered by Hadley and Hayward, specifically due to the much smaller explanatory distance between the architecture and the underlying scientific theories invoked contextually as part of the explanation. In view of the above arguments, the S11 architecture is thus capable of answering the challenge, in terms of the variant form of the criticism proposed by Aizawa (1997a, 1997b) as discussed in section 2.8.1, for a bona fide scientific explanation of the aforementioned systematicity properties in humans (as promised in chapter 3).

7.2 Summary of Contributions

A novel hybrid connectionist architecture (S11) has been developed in this thesis. As described in section 7.1, the architecture is experimentally shown to exhibit strong systematicity, exhibit partial semantic systematicity, and thus is behaviourally adequate with respect to Syntax 1 and 2. Further, the S11 architecture is psychologically plausible to some degree, but certainly more so than the connectionist architectures surveyed in sections 2.3 and 2.7 (especially including, as a point of comparison, the architecture by Hadley & Hayward, 1997). The S11 architecture is also oriented towards increasing neurobiological plausibility.

As noted in section 7.1, the S11 architecture provides a possible bona fide *scientific*

explanation (and not just an exhibition) of the phenomenon of linguistic semantic systematicity in humans, and thus satisfies the challenge of the variant form of the criticism proposed by Aizawa (1997a, 1997b) as discussed in section 2.8.1. Namely, language processing is semantically systematic mainly because human *conceptualization* of the external world is systematic, compositional, and any fragment of conceptualization can be identified and manipulated independently and in isolation from all other fragments.

Furthermore, the experimentally shown behavioural adequacy of the developed architecture provides confirmatory evidence that corroborates, to some degree, the proposed S11 architecture and conceptual semantics, largely as proposed in Cognitive Grammar. This means, in summary, that the evidence suggests understanding (1) that language meaning is in part holistic; (2) that meaning is in part also conceptual, systematic, and structured in ways similar to how it is analyzed in Cognitive Grammar; (3) that language comprehension in part need not require explicit syntactic mental structures of the variety advanced by generative linguists; and (4) that the ability to identify common goal objects may very well play a large role in language learning. Therefore, the S11 architecture demonstrates one possible way of providing a processing model of conceptual semantics, largely as proposed in Cognitive Grammar, that can be implemented within a connectionist framework.

7.3 Future Research

Recall that the general goal that started off this thesis is to explore and deepen, using machine models within a naturalistic scientific framework, our understanding of human cognition, and more specifically, the human cognitive ability to comprehend language. Therefore, although increasing behavioural adequacy is desirable, it is imperative that future research is done to enhance the cognitive, psychological, and neurobiological plausibility of the architecture proposed in this thesis.

In particular, recall that in this thesis, the S11 architecture constitutes an incomplete model, as the features included in the scene structure are fully manifested, static, and atemporal. In future work, a dynamic and real-time model of conceptual cognitive structure would have to be developed.

The scene structure is assumed to be itself a result of connectionistic processing of experiential perceptual input. This thesis did not elaborate on how that is possible, except to suggest that Deep Belief Nets might be a possible architecture for that kind of processing.

Future research should address this point.

Although the software implementation of the S11 architecture in this thesis is technically a hybrid connectionist model, it has been argued in this thesis that, in principal, it is reducible to a non-hybrid connectionist model. Therefore, future research should address the reduction of the S11 architecture into a fully non-hybrid model.

As explained in section 7.1, linguistic systematicity is probably a result of human *conceptualization* of the external world being systematic and compositional, such that any fragment of conceptualization can be identified and manipulated independently and in isolation from all other fragments. In turn, conceptualization is itself systematic probably because the external world is also systematic in some sense, and the perceptual modules in our brain processes perception into systematic conceptualizations. How the external world is systematic in any sense is a question for physics. As for how the perceptual modules manage to produce systematic conceptualizations within a connectionistic network, that is a question that will have to be explored in future work that deals with processing of perceptual input.

It was noted in section 4.3 that the brain will either (1) implement multiple areas of storage such that each can be modelled by a scene structure, or (2) implement multiple areas of storage such that each can be modelled by a plate, and such that an underlying neurobiological processing mechanism exists that can be modelled as the numerous plates being reorganized into multiple scene structures. In either case, each scene structure would model the storage of just one construal or aspect of the conceptualization of a complete situation, such that the situation may be conceptualized from currently occurring perceptual stimulations or be conceptualized from what would phenomenologically be described as vivid imagination. This complete conceptualization is referred to as the conceptual substrate by Langacker (2008, p. 54). This thesis only deals with a single scene structure and does not model the entire conceptual substrate, a task best left to future research.

References

Aizawa, K. K. (1997a, February). Exhibiting versus Explaining Systematicity: A Reply to Hadley and Hayward. *Minds and Machines*, *7*(1), 39–55.

Aizawa, K. K. (1997b, June). Explaining Systematicity. *Mind & Language*, *12*(2), 115–136.

Aizawa, K. K. (2003). *The Systematicity Arguments* (1st ed.). Springer. Paperback.

Aydede, M. (1997). Language of Thought: The Connectionist Contribution. *Minds and Machines*, *7*(1), 57–101.

Bengio, Y. (2007). *Learning deep architectures for AI* (Tech. Rep.). Dept. IRO, Universite de Montreal.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning* (Corrected at Sixth Printed ed.). New York, NY: Springer. Hardcover.

Brakel, P., & Frank, S. L. (2009). Strong systematicity in sentence processing by simple recurrent networks. In N. Taatgen & H. van Rijn (Eds.), *Cogsci 2009 proceedings.* Austin, TX. Web site: http://csjarchive.cogsci.rpi.edu/proceedings/2009/papers/344/index.html.

Davidson, D. (1967, January). Truth and Meaning. *Synthese*, *17*(1), 304–323.

Dawson, M. R. W., & Shamanski, K. S. (1994). Connectionism, Confusion, and Cognitive Science. *The Journal of Intelligent Systems*, *4*, 215–262.

Dehaene-Lambertz, G., Hertz-Pannier, L., & Dubois, J. (2006, July). Nature and nurture in language acquisition: anatomical and functional brain-imaging studies in infants. *Trends in Neurosciences*, *29*(7), 367–373.

Dennett, D. C. (1987a). *The Intentional Stance*. The MIT Press. Paperback.

Dennett, D. C. (1987b). Three Kinds of Intentional Psychology. In *The intentional stance* (p. 43+). The MIT Press. Paperback.

Dennett, D. C. (1991). Mother Nature vs. the Walking Encyclopaedia. In W. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory.* Hillsdale, NJ: Erlbaum.

Ding, Y., & Marotte, L. R. (1997, August). Retinotopic order in the optic nerve and superior colliculus during development of the retinocollicular projection in the wallaby (Macropus eugenii). *Anatomy and embryology*, *196*(2), 141–158.

Elman, J. L. (1990, June). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1991, September). Distributed Representations, Simple Recurrent Networks, And Grammatical Structure. *Machine Learning*, *7*(2), 195–225.

Fiez, J. A. (1997). Phonology, semantics, and the role of the left inferior prefrontal cortex. *Human brain mapping*, *5*(2), 79–83.

Filley, C. M. (2010, June). White Matter: Organization and Functional Relevance. *Neuropsychology Review*, *20*(2), 158–173.

Fodor, J. A. (1978). Tom Swift and his procedural grandmother. *Cognition*, *6*(3), 229–247.

Fodor, J. A. (1999, July). Information and Representation. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (pp. 513–524). Cambridge, Massachusetts: The MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988, March). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Frank, S. L., Haselager, W. F. G., & van Rooij, I. (2009, March). Connectionist Semantic Systematicity. *Cognition*, *110*(3), 358–379.

Frege, G. (2000, November). On Sense and Nominatum. In A. P. Martinich (Ed.), *The philosophy of language* (Fourth ed., pp. 199–211). USA: Oxford University Press. Paperback.

Gisiger, T., & Boukadoum, M. (2011). Mechanisms Gating the Flow of Information in the Cortex: What They Might Look Like and What Their Uses may be. *Frontiers in computational neuroscience*, *5*.

Grice, H. P. (1975, June). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, volume 3: Speech acts* (Vol. 3, pp. 41–58).

Grossberg, S., & Vladusich, T. (2010, October). How do children learn to follow gaze, share joint attention, imitate their teachers, and use tools during social interactions? *Neural Networks*, *23*(8-9), 940–965.

Hadley, R. F. (1989, March). A default-oriented theory of procedural semantics. *Cognitive Science*, *13*(1), 107–137.

Hadley, R. F. (1994, September). Systematicity in Connectionist Language Learning. *Mind & Language*, *9*(3), 247–272.

Hadley, R. F. (1997, November). Explaining Systematicity: A Reply to Kenneth Aizawa. *Minds and Machines*, *7*(4), 571–579.

Hadley, R. F. (1999). Connectionism and Novel Combinations of Skills: Implications for Cognitive Architecture. *Minds and Machines*, *9*(2), 197–221.

Hadley, R. F. (2000). Cognition and the computational power of connectionist networks. *Connection Science*, *12*(2), 95–110.

Hadley, R. F. (2009). The Essential Role of Consciousness in Mathematical Cognition. In N. Taatgen & H. van Rijn (Eds.), *Cogsci 2009 proceedings.*

Hadley, R. F. (2011, March). *Binding Concepts Together: How Does the Brain Do It?* Defining Cognitive Science Speaker Series.

Hadley, R. F., & Cardei, V. C. (1999, March). Language acquisition from sparse input without error feedback. *Neural Networks*, *12*(2), 217–235.

Hadley, R. F., & Hayward, M. B. (1997, February). Strong Semantic Systematicity from Hebbian Connectionist Learning. *Minds and Machines*, *7*(1), 1–37.

Hadley, R. F., Rotaru-Varga, A., Arnold, D. V., & Cardei, V. C. (2001). Syntactic systematicity arising from semantic predictions in a Hebbian-competitive network. *Connection Science*, *13*(1), 73–94.

Hagler, D. J., & Sereno, M. I. (2006, January). Spatial maps in frontal and prefrontal cortex. *NeuroImage*, *29*(2), 567–577.

Hardwig, J. (1991, December). The Role of Trust in Knowledge. *The Journal of Philosophy*, *88*(12), 693–708.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (Second ed.). Springer.

Herculano-Houzel, S., Mota, B., Wong, P., & Kaas, J. H. (2010, November). Connectivity-driven white matter scaling and folding in primate cerebral cortex. *Proceedings of the National Academy of Sciences*, *107*(44), 19008–19013.

Hinton, G. E. (2010, January). Learning to represent visual input. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1537), 177–184.

Hsiao, J. H. (2002). *Dealing with Semantic Anomalies in a Connectionist Network for Word Prediction*. Master's thesis, School of Computing Science, Simon Fraser University.

Hyötyniemi, H. (1996, August). Turing Machines are Recurrent Neural Networks. In J. Alander, T. Honkela, & M. Jakobsson (Eds.), *Proceedings of step'96.* Finnish Artificial Intelligence Society. Electronically.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford University Press, USA. Hardcover.

Jackson, S. A. (1996). *Connectionism and Meaning: From Truth Conditions to Weight Representations*. Ablex Publishing Corporation. Paperback.

Johnson, M., Demuth, K., Frank, M., & Jones, B. (2010). Synergies in learning words and their referents. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems 23* (pp. 1018–1026).

Kaas, J. H. (1997). Topographic maps are fundamental to sensory processing. *Brain research bulletin*, *44*(2), 107–112.

Katz, J. J., & Postal, P. M. (1978). *An Integrated Theory of Linguistic Descriptions*. The MIT Press. Paperback.

Katz, P. S. (2003, July). Synaptic Gating: The Potential to Open Closed Doors. *Current Biology*, *13*(14), R554–R556.

Kim, J. (2005). *Philosophy of Mind* (Second ed.). Westview Press. Paperback.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd ed.). University Of Chicago Press. Paperback.

Kunes, S., Wilson, C., & Steller, H. (1993, February). Independent guidance of retinal axons in the developing visual system of Drosophila. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *13*(2), 752–767.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things* (1987th ed.). University Of Chicago Press. Paperback.

Langacker, R. W. (1991). *Foundations of Cognitive Grammar: Volume II: Descriptive*

Application (1st ed.). Stanford University Press. Hardcover.

Langacker, R. W. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford University Press, USA. Paperback.

LeCun, Y., Bottou, L., Orr, G., & Müller, K. (1998, March). Efficient BackProp. In G. Orr & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade* (Vol. 1524, p. 546). Berlin, Heidelberg: Springer Berlin / Heidelberg.

Lewis, D. (1970, December). General semantics. *Synthese*, *22*(1), 18–67.

Lit, L., Schweitzer, J., & Oberbauer, A. (2011, January). Handler beliefs affect scent detection dog outcomes. *Animal Cognition*, 1–8.

Mailly, P., Charpier, S., Mahon, S., Menetrey, A., Thierry, A. M., Glowinski, J., et al. (2001, September). Dendritic Arborizations of the Rat Substantia Nigra Pars Reticulata Neurons: Spatial Organization and Relation to the Lamellar Compartmentation of Striato-Nigral Projections. *The Journal of Neuroscience*, *21*(17), 6874–6888.

Marcus, G. F. (1998, December). Rethinking Eliminative Connectionism. *Cognitive Psychology*, *37*(3), 243–282.

Marcus, G. F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. The MIT Press. Hardcover.

McLaughlin, T., & O'Leary, D. D. M. (2005). Molecular gradients and development of retinotopic maps. *Annual review of neuroscience*, *28*(1), 327–355.

Memmi, D. (1990, April). Connectionism and artificial intelligence as cognitive models. *AI & Society*, *4*(2), 115–136.

Midgley, M. (1994). *Science as Salvation: A Modern Myth and its Meaning* (1st ed.). Routledge. Paperback.

Montague, R. (1973). The Proper Treatment of Quantification in Ordinary English. , 221–242.

Moreira, M., & Fiesler, E. (1995, October). *Neural Networks with Adaptive Learning Rate and Momentum Terms* (Tech. Rep.). Martigny, Valais, Suisse: Institut Dalle Molle D'Intelligence Artificielle Perceptive.

Morel, A., Liu, J., Wannier, T., Jeanmonod, D., & Rouiller, E. M. (2005, February). Divergence and convergence of thalamocortical projections to premotor and supplementary motor cortex: a multiple tracing study in the macaque monkey. *European Journal of Neuroscience*, *21*(4), 1007–1029.

Movellan, J. R. (1991). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, & G. E. Hinton (Eds.), *Connectionist models: Proceedings of the 1990 summer school* (pp. 10–17). San Mateo, California: Morgan Kaufmann Publishers, Inc.

Nagel, T. (2002, March). Knowledge. In K. B. Wray (Ed.), *Knowledge & Inquiry* (p. 205+). Broadview Press.

Neumann, J. (2002, June). Learning the systematic transformation of holographic reduced representations. *Cognitive Systems Research*, *3*(2), 227–235.

O'Donnell, P., & Grace, A. A. (1995, May). Synaptic interactions among excitatory afferents to nucleus accumbens neurons: hippocampal gating of prefrontal cortical input. *The*

Journal of Neuroscience, *15*(5), 3622–3639.

Oldham, M. C., & Geschwind, D. H. (2005, March). Evolutionary Genetics: The human brain — adaptation at many levels. *European Journal of Human Genetics*, *13*(5), 520–522.

O'Leary, D. D. M., Yates, P. A., & McLaughlin, T. (1999, January). Molecular Development of Sensory Maps. *Cell*, *96*(2), 255–269.

Partee, B. (1973, October). Some transformational extensions of Montague grammar. *Journal of Philosophical Logic*, *2*(4), 509–534.

Pfungst, O. (1911). *Clever Hans (The Horse of Mr. von Osten): A Contribution to Experimental Animal and Human Psychology* (C. L. Rahn, Trans.). New York: Henry Holt and Company.

Pulvermüller, F. (2005, June). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(7), 576–582.

Quine, W. V. (1960). *Word and Object*. The MIT Press. Paperback.

Quine, W. V. (1987, January). Indeterminacy of Translation Again. *The Journal of Philosophy*, *84*(1).

Ramsey, W., Stich, S., & Garon, J. (1990). Connectionism, Eliminativism and The Future of Folk Psychology. *Philosophical Perspectives*, *4*, 499–533.

Raymond, E. S. (Ed.). (2003, December). *The Jargon File (version 4.4.7).* Online. Available from `http://www.catb.org/jargon/html/koans.html#id3141241`

Rockland, K. S., & Drash, G. W. (1996). Collateralized divergent feedback connections that target multiple cortical areas. *Journal of Comparative Neurology*, *373*(4), 529–548.

Rumelhart, D. E., & McClelland, J. L. (1986). PDP Models and General Issues in Cognitive Science. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Foundations* (Vol. 1, pp. 110–146). Cambridge, MA, USA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Foundations* (Vol. 1, pp. 151–193). Cambridge, MA, USA: MIT Press.

Saleem, K. S., & Tanaka, K. (1996, August). Divergent Projections from the Anterior Inferotemporal Area TE to the Perirhinal and Entorhinal Cortices in the Macaque Monkey. *The Journal of Neuroscience*, *16*(15), 4757–4775.

Schaffner, K. F., Salmon, W. C., Norton, J. D., McGuire, J. E., Machamer, P., & Lennox, J. G. (1999). *Introduction to the Philosophy of Science* (1st ed.). Hackett Pub Co Inc. Paperback.

Schulte, O. (n.d.). *My View Of Philosophy.* Retrieved 2 July 2011 from web page. Available from `http://www.cs.sfu.ca/~oschulte/phil.html`

Sereno, M. I. (1991). Language and the primate brain. In *Proceedings, thirteenth annual conference of the cognitive science society* (pp. 79–84). Lawrence Erlbaum Associates.

Siegelmann, H. T., & Margenstern, M. (1999, June). Nine switch-affine neurons suffice for turing universality. *Neural Networks*, *12*, 593–600.

Smolensky, P., & Legendre, G. (2006). Principles of the Integrated Connectionist / Symbolic Cognitive Architecture. In P. Smolensky & G. Legendre (Eds.), *The harmonic mind: From neural computation to optimality-theoretic grammar* (Vol. 1, pp. 63–97). Cambridge, MA: MIT Press.

Soames, S. (1992, February). Truth, meaning, and understanding. *Philosophical Studies*, *65*(1), 17–35.

Solomon, S. G., & Lennie, P. (2007, April). The machinery of colour vision. *Nature Reviews Neuroscience*, *8*(4), 276–286.

Stewart, T., & Eliasmith, C. (2009). Compositionality and Biologically Plausible Models. In M. Werning, W. Hinzen, & E. Machery (Eds.), *Oxford handbook of compositionality.* Oxford University Press.

Tamamaki, N., & Nojyo, Y. (1995). Preservation of topography in the connections between the subiculum, field CA1, and the entorhinal cortex in rats. *Journal of Comparative Neurology*, *353*(3), 379–390.

Thagard, P. (2010, June). *Cognitive Science.* Web page. Stanford, CA 94305. Available from `http://plato.stanford.edu/entries/cognitive-science/`

Thivierge, J.-P., & Marcus, G. F. (2007, June). The topographic brain: from neural connectivity to cognition. *Trends in Neurosciences*, *30*(6), 251–259.

Tong, M. H., Bickett, A. D., Christiansen, E. M., & Cottrell, G. W. (2007, April). Learning grammatical structure with Echo State Networks. *Neural Networks*, *20*(3), 424–432.

Tootell, R. B., Switkes, E., Silverman, M. S., & Hamilton, S. L. (1988, May). Functional anatomy of macaque striate cortex. II. Retinotopic organization. *The Journal of Neuroscience*, *8*(5), 1531–1568.

van Gelder, T. (1990, September). Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, *14*(3), 355–384.

van Gelder, T. (1993, October). Connectionism and the mind-body problem: exposing the distinction between mind and cognition. *Artificial Intelligence Review*, *7*(5), 355–369.

van Inwagen, P. (1993). *Metaphysics* (First Edition ed.; N. Daniels & K. Lehrer, Eds.). Boulder: Westview Press. Paperback.

Verhage, M., Maia, A. S., Plomp, J. J., Brussaard, A. B., Heeroma, J. H., Vermeer, H., et al. (2000, February). Synaptic Assembly of the Brain in the Absence of Neurotransmitter Secretion. *Science*, *287*(5454), 864–869.

Vermazen, B. (1967, January). Review of Jerrold Katz and Paul Postal, *an integrated theory of linguistic descriptions*, and Katz, *philosophy of language*. *Synthese*, *17*(1), 350–365.

Weisz, N., Wienbruch, C., Hoffmeister, S., & Elbert, T. (2004, May). Tonotopic organization of the human auditory cortex probed with frequency-modulated tones. *Hearing Research*, *191*(1-2), 49–58.

Woods, W. A. (1981). Procedural Semantics as a Theory of Meaning. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 300–334). Cambridge, UK: Cambridge University Press.

Xie, X., & Seung, H. S. (2003). Equivalence of Backpropagation and Contrastive Hebbian

Learning in a Layered Network. *Neural Computation*, *15*, 441–454.

Zhang, K., & Sejnowski, T. J. (2000, May). A universal scaling law between gray matter and white matter of cerebral cortex. *Proceedings of the National Academy of Sciences*, *97*(10), 5621–5626.

Appendix A

Entity Features

In the following tables, an entry of "1" denotes the feature for that row, as indicated in the left-most column, has been assigned to the entity for that column, as indicated in the column heading. Likewise, a blank entry indicates the absence of that feature being assigned to that entity. See also section 4.4.

Table A.1: Features for thing entities (see also table A.2)

	WOMEN	BB-BATS	BOYS	TABLES	CHAIRS	BIRDS
:physical	1	1	1	1	1	1
:inanimate		1		1	1	
:animate	1		1			1
:has-color				1	1	1
:green					1	1
:blue				1		1
:has-weight	1	1	1	1	1	1
:heavy				1		
:light		1	1			1
:medium-heavy	1				1	
:has-height	1		1	1	1	1
:short			1		1	1
:medium-tall	1			1		
:has-shape	1	1	1	1	1	1
:rigid		1		1	1	
:flexible	1		1			1

Table A.1: (continued)

	WOMEN	BB-BATS	BOYS	TABLES	CHAIRS	BIRDS
:tubular		1				
:round						1
:has-texture	1	1	1	1	1	1
:smooth	1	1		1	1	
:hairy			1			1
:has-size	1	1	1	1	1	1
:small		1	1			1
:large				1		
:medium-large	1				1	
:has-behavior	1		1			1
:bites						1
:emotive	1					
:has-face	1		1			1
:flat-face			1			
:long-snout						1
:small-nose	1					
:has-legs	1		1			1
:two-legs	1		1			1
:make-sounds	1		1			1
:squeaks						1
:talks	1		1			
:laughs	1		1			
:has-location-horizontal	1	1	1	1	1	1
:left-of-stage	1					1
:center-stage		1		1		
:right-of-stage			1		1	
:has-location-vertical	1	1	1	1	1	1
:far-above-stage						1
:on-stage	1	1	1	1	1	
:has-distance	1	1	1	1	1	1
:far-from-speaker						1
:medium-far-from-speaker		1		1	1	
:near-speaker			1			
:close-by-speaker	1					

Table A.2: Features for thing entities (see also table A.1)

	GIRLS	DOGS	CATS	MICE	MEN	BASEBALLS
:physical	1	1	1	1	1	1
:inanimate						1
:animate	1	1	1	1	1	
:has-weight	1	1	1	1	1	1
:heavy					1	
:light	1		1	1		1
:medium-heavy		1				
:has-height	1	1	1	1	1	
:tall					1	
:short	1	1	1	1		
:has-shape	1	1	1	1	1	1
:rigid						1
:flexible	1	1	1	1	1	
:round				1		1
:has-texture	1	1	1	1	1	1
:furry		1	1			
:smooth	1					1
:hairy				1	1	
:has-size	1	1	1	1	1	1
:small	1		1	1		1
:large					1	
:medium-large		1				
:has-behavior	1	1	1	1	1	
:bites		1	1	1		
:emotive	1					
:has-face	1	1	1	1	1	
:flat-face			1		1	
:long-snout		1		1		
:small-nose	1		1			
:has-legs	1	1	1	1	1	
:two-legs	1				1	
:four-legs		1	1	1		
:make-sounds	1	1	1	1	1	
:barks		1				

Table A.2: (continued)

	GIRLS	DOGS	CATS	MICE	MEN	BASEBALLS
:meows			1			
:squeaks				1		
:talks	1				1	
:laughs	1				1	
:has-location-horizontal	1	1	1	1	1	1
:left-of-stage		1			1	
:center-stage						1
:right-of-stage	1		1	1		
:has-location-vertical	1	1	1	1	1	1
:on-stage	1	1	1	1	1	1
:has-distance	1	1	1	1	1	1
:far-from-speaker				1		
:medium-far-from-speaker		1	1			1
:near-speaker	1					
:close-by-speaker					1	

Table A.3: Features for process relationship entities (see also table A.4)

	LOVE	CHASE	SEES	SWING	FOLLOW	BUMP
:physical-motion		1		1	1	1
:rapid-motion		1		1		
:medium-rapid-motion					1	1
:smooth-motion				1	1	
:jumpy-motion		1				1
:involves-animate	1	1	1	1	1	1
:involves-inanimate				1	1	1
:involves-perceiver	1		1			
:feeling-nice	1					
:is-occurring	1	1	1	1	1	1
:level-to-ground		1			1	
:moving-nearer		1			1	1
:moving-in-sync	1					
:closely-touching	1			1		1

Table A.3: (continued)

	LOVE	CHASE	SEES	SWING	FOLLOW	BUMP
:separated-from		1			1	
:separated-far-from			1			
:occurring-together	1	1	1	1	1	1
:path	1	1	1	1	1	1

Table A.4: Features for process relationship entities (see also table A.3)

	HIT	AVOID	LIKE	CONSUME	DISLIKE
:physical-motion	1	1		1	
:rapid-motion	1				
:slow-motion		1		1	
:jumpy-motion	1				
:involves-animate	1	1	1	1	1
:involves-inanimate	1		1		1
:involves-perceiver		1	1	1	1
:feeling-bad		1			1
:feeling-nice			1	1	
:is-occurring	1	1	1	1	1
:level-to-ground		1			
:moving-nearer	1			1	
:moving-away		1			1
:moving-in-sync			1		
:close-together			1		
:closely-touching	1			1	
:separated-far-from					1
:occurring-together	1	1	1	1	1
:path	1	1	1	1	1

Table A.5: Features for non-processual relationship entities that has as its trajector a relationship (i.e., what are ordinarily expressed by adverbs)

	PHYSICALLY	RAPIDLY	SLOWLY	SMOOTHLY	JUMPILY	NICELY	BADLY
:physical	1						
:heavy			1				
:light		1					
:smooth				1			
:rough					1		
:has-behavior	1	1	1	1	1	1	1
:emotive						1	1
:physical-motion	1	1	1	1	1		
:rapid-motion		1					
:slow-motion			1				
:smooth-motion				1			
:jumpy-motion					1		
:involves-perceiver						1	1
:feeling-bad							1
:feeling-nice						1	

Table A.6: Features for non-processual relationship entities that has as its trajector a thing (i.e., what are ordinarily expressed by adjectives. See also table A.7)

	RED	GREEN	BLUE	HEAVY	LIGHT
:physical	1	1	1	1	1
:inanimate	1	1	1	1	1
:has-color	1	1	1		
:red	1				
:green		1			
:blue			1		
:has-weight				1	1
:heavy				1	
:light					1

Table A.7: Features for non-processual relationship entities that has as its trajector a thing (i.e., what are ordinarily expressed by adjectives. See also table A.6)

	TALL	SHORT	FURRY	SMOOTH	HAIRY
:physical	1	1	1	1	1
:inanimate	1	1	1	1	1
:has-height	1	1			
:tall	1				
:short		1			
:flexible			1		
:has-texture				1	
:furry			1		1
:smooth				1	
:hairy					1

Table A.8: Features for entities ordinarily expressed by subordinators, and the START entity

	WITH	FROM	THAT	START
:far-from-speaker			1	
:moving-away		1		
:moving-in-sync	1			
:separated-from	1	1		
:separated-far-from			1	
:occurred-before		1	1	
:occurring-together	1			
:focal-prominence-low	1	1	1	1
:path	1	1	1	
:objectivity-high			1	1
:objectivity-medium	1	1		

Index

Author Index