Infinite Languages, Finite Minds

Connectionism, Learning and Linguistic Structure

Morten Hyllekvist Christiansen

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
to the
University of Edinburgh
1994

Abstract

This thesis presents a connectionist theory of how infinite languages may fit within finite minds. Arguments are presented against the distinction between linguistic competence and observable language performance. It is suggested that certain kinds of finite state automata—i.e., recurrent neural networks—are likely to have sufficient computational power, and the necessary generalization capability, to serve as models for the processing and acquisition of linguistic structure. These arguments are further corroborated by a number of computer simulations, demonstrating that recurrent connectionist models are able to learn complex recursive regularities and have powerful generalization abilities. Importantly, the performance evinced by the networks are comparable with observed human behavior on similar aspects of language. Moreover, an evolutionary account is provided, advocating a learning and processing based explanation of the origin and subsequent phylogenetic development of language. This view construes language as a nonobligate symbiant, arguing that language has evolved to fit human learning and processing mechanisms, rather than *vice versa*. As such, this perspective promises to explain linguistic universals in functional terms, and motivates an account of language acquisition which incorporates innate, but *not* language-specific constraints on the learning process. The purported poverty of the stimulus is re-appraised in this light, and it is concluded that linguistic structure may be learnable by bottom-up statistical learning models, such as, connectionist neural networks.

Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

December 1994

Acknowledgements

This thesis reflects my work on language learning and processing within a connectionist framework. I have worked on these issues for more than four years. After being introduced to cognitive science at the University of Warwick in 1989 as a MSc student in the department of Psychology, I was fortunate to study for my PhD at the Centre for Cognitive Science, the University of Edinburgh. Without the outstanding resources of this department—human as well as technical—this thesis would never have come into being.

The research presented here has benefitted from discussions with a number of people who kindly have shared their time and knowledge with me. First and foremost, I would like to thank my two supervisors Nick Chater and Paul Schweizer. They have been very supportive and helpful during my work on this thesis. I am especially grateful to Nick with whom I have co-authored several papers. It is my hope that the future may bring many more such collaborations. I am indebted to Andy Clark, Eric Dietrich, Jeff Elman, and Heather Morrison for commenting on early versions of thesis chapters. Many others deserve thanks as well. Adele Abrahamsen, Bill Bechtel, Keith Butler, Dave Chalmers, Bob Hadley, and Mark Rollins have contributed with their comments and suggestions on my ideas, as have the students in my graduate course in connectionist modeling and natural language processing at Washington University in St. Louis. At Indiana University, Bob French, Mike Gasser, Rob Goldstone, Doug Hofstadter, Gary McGraw, Jim Marshall, and Bob Port also helped shape this thesis. Joe Devlin, Martin Pickering, Jenny Saffran, and Mitch Sommers helped me track down some important references. Paul Cairns provided invaluable 'on-line' help regarding the *Xerion* simulator used in chapter 3 (distributed by the University of Toronto). Jeff Elman kindly made his simulator, *Tlearn*, available for my use in the simulations presented in chapter 4, and provided advice regarding my experiments. Thanks are also due to Helga Keller for her amicable secretarial support.

My thesis work would not have been possible without support from several institutions. The Danish Research Academy funded my first three years of study. Knud

Højgaards Fond twice awarded me funds for study visits in the U.S.A. Doug Hofstadter generously provided me with office space and a computer in the Center for Research on Concepts and Cognition at Indiana University during a two-months visit in July/August 1992, and later awarded me with a four-month research assistantship during the spring of 1993. A McDonnell Postdoctoral Fellowship in the Philosophy-Neuroscience-Psychology Program at Washington University in St. Louis supported me during the final stages of my research. Jeff Elman was very helpful during my short visit to the Center for Research in Language at the University of California, San Diego, in the summer of 1994. Finally, my attendance as a Fellow at the McDonnell Summer Institute in Cognitive Neuroscience, July 1994, provided some last minute inspirations.

And of course, my most heartfelt gratitude and my undivided love go to Anita. You have made the greatest effort to make the writing process go as smoothly as possible. You have provided excellent help with all aspects of my thesis writing. You have patiently listened to me going on about every little detail in my thesis, spent hours reading and commenting on my chapters, correcting mistakes that nobody else found. And you dragged me away from the computer when I needed it the most. Without you and my cat Betty, I am not sure I would have survived the final stages of thesis writing. In short: Anita, du gør livet værd at leve!

Contents

Chapter 1

Introduction

Imagine a world without language. What would be the fate of humanity, if language as we know it suddenly ceased to exist? Would we end up as the ill-fated grunting hominids in "Planet of the Apes"? Not likely, perhaps, but try to imagine the devastating consequences of language loss on a more personal scale. Language is so tightly interwoven into the very fabric of our everyday lives that losing even parts of it would have far-reaching detrimental repercussions. Consider, for example, the problems facing someone with agrammatic aphasia as evidenced in the following speech sample from a patient explaining that he has returned to the hospital for some dental work:

> Ah... Monday... ah, Dad and Paul Haney [*referring to himself by his full name*] and Dad... hospital. Two... ah, doctors..., and ah... thirty minutes... and yes... ah... hospital. And, er, Wednesday... nine o'clock. And er Thursday, ten o'clock... doctors. Two doctors... and ah... teeth. Yeah,... fine. (Ellis & Young, 1988: p. 242)

Now, imagine an evolutionary scenario in which the hominids did not evolve a capability for language. In such a picture, it is obvious that humankind would never have ended up where we are today. For example, instructing the young in new skills without the use of language would have been difficult, once we get beyond a certain skill complexity. The development and spread of new technologies would therefore be severely impeded, and we might never have evolved beyond being hunters and gatherers.

As should be clear, language is a very powerful means of communication, apparently unique to humans, allowing us to communicate about an unbounded number of different objects, situations and events. Language permits us to transfer cultural information readily from generation to generation—originally, in terms of verbal instructions and enlightening tales; later, writing ensured a more constant source of storing information, making it easier to share knowledge across time; and most recently, computers have

allowed us to communicate rapidly over great distances (for instance, via the various INTERNET facilities, such as, email and Mosaic). The main reason why language is such a powerful means of communication is—as first pointed out by Wilhelm von Humboldt (quoted in Chomsky, 1965: p. v)—that it "makes infinite use of finite means". That is, we can use language to describe anything that is within the limits of our perceptual and intellectual capacities. In other words, despite the finite resources of the human mind, we are able to produce languages that are infinite in nature.

As one of the hallmarks of human cognition, language has received much attention within the science dedicated to the study of the human mind: *cognitive science*. For many years this study was dominated by the 'computer metaphor of the mind', proposing an analogy between the processing of symbols in a digital computer and the workings of the mind. More recently, a different view of cognitive processing has emerged, based on artificial neural networks of inter-connected, very simple processing units. The two different approaches are often construed as two different paradigms (e.g., Schneider, 1987), and—as was to be expected (cf. Kuhn, 1972)—the proliferation of connectionism in the second half of the 80's led to much subsequent debate (e.g., Chalmers, 1990b; Chater & Oaksford, 1990; Christiansen & Chater, 1992, 1994; Fodor & Pylyshyn, 1988; Fodor & McLaughlin, 1990; Hadley, 1994a, 1994b; Niklasson & Sharkey, 1992; Niklasson & van Gelder, 1994; Smolensky, 1987, 1988; van Gelder, 1990). In this debate, language has been considered by many to be the largest stumbling block for connectionism. Indeed, Pinker & Prince (1988) argue:

> From its inception, the study of language within the framework of generative grammar has been a prototypical example of how fundamental properties of a cognitive domain can be explained within the symbolic paradigm. ...Language has been the domain most demanding of articulated symbol structures governed by rules and principles and it is also the domain where such structures have been explored in the greatest depth and sophistication ...Many observers thus feel that connectionism, as a radical restructuring of cognitive theory, will stand or fall depending on its ability to account for human language. (p. 78)

More recently, Chomsky (1993) has provided a rather negative assessment of connectionism's current contribution to our general understanding of cognition and, in particular, of language.

> Connectionism is a radical abstraction from what's known about the brain and the brain sciences ...There's no reason to believe you're abstracting the right thing. There's no evidence for it. In the case of language, the evidence for connectionist models is, for the moment, about zero. The most trivial

> problems that have been addressed—like learning a few hundred words—have been total failures." (p. 85–86)

In this thesis, I take up the challenge posed by Pinker & Prince, showing how a connectionist picture might explain our infinite use of language given our finite minds, and thereby rebut Chomsky's negative assessment of connectionism. My account involves a view of language processing, which appear to be more in line with psycholinguistic data, as well as a learning and processing based perspective on the acquisition and evolution of language. But before outlining this theory, I will make some brief methodological remarks.

1.1 Some Methodological Preliminaries

In discussions of symbolic and connectionist approaches to cognitive science, the historical predominance of the symbolic view has meant that, to some extent at least, the ground rules concerning what key cognitive phenomena must be explained and what counts as a good explanation, have been set in largely symbolic terms (van Gelder, 1991). Thus, if connectionism amounts to a genuinely new paradigm for the understanding of mind, there is a very real danger of falling into what I elsewhere (Christiansen & Chater, 1992) have called the *'incommensurability trap'*. That is, connectionist models may be unfairly judged either because they fail to fit the classical standards or because when they are made to fit, the resulting explanation looks forced and unattractive. The danger is analogous to that of judging vegetarian food by the standards of the butcher. After all, connectionism—construed as a new paradigm (e.g., Schneider, 1987)—may involve a revolutionary reconstruction of the field from new fundamentals, leading to changes in methodology and basic theoretical assumptions. Since rival paradigms prescribe different sets of standards and principles, connectionist and classical approaches to cognitive science may also differ on what constitute meaningful and legitimate scientific questions. Due to this incommensurability, discussions between proponents of different paradigms on the issue of paradigm choice often become circular. Each group will tend to praise their own paradigm and criticize the other's with arguments based on their own paradigm. In other words, when comparing and assessing the individual explanatory power of rival paradigms, the incommensurability trap constitutes a nontrivial methodological obstacle to negotiate since it involves engaging in the process of *radical translation* (Quine, 1960). Or so, much philosophy of science would have us to believe (e.g., Kuhn, 1970). In any case, there are signs that communication is becoming difficult, and hence it is imperative that the merits of connectionism are judged from "within"—i.e., on its own terms—not through the

looking glass of the classical paradigm[1].

The incommensurability trap might manifest itself in different ways. It is often presupposed that the level of explanation within cognitive science must necessarily be that of the classical paradigm. This is, for example evident when Prince & Pinker (1989), in their criticism of the Rumelhart & McClelland (1986) model of the acquisition of English past tense, assert that "neuroscientists study firing rates, excitation, inhibition, plasticity; cognitive scientists study *rules, representations, symbol systems*" (p. 1: my emphasis). If cognition has to be couched in terms of 'rules', 'symbols', and so on, then connectionism is likely to fail as a genuine new approach to cognitive science. However, these terms are theoretical constructs belonging to the classical paradigm (at least, in their most typical instantiations). As such, it is fallacious to assume that such classical constructs are a necessary part of cognitive explanations. That connectionist models might not be able to embody rules and context-independent symbols should therefore not be taken *a priori* as evidence of shortcomings. Instead, connectionism must be judged on whether they can account for the (uninterpreted) data that originally let to the postulation of those classical constructs.

Another consequence of the incommensurability trap is the tendency to hold connectionist models to a higher standard than similar symbolic models. Often a particular connectionist model is criticized for not reaching a certain level of performance—as when Hadley (1994a) criticizes connectionist models of language acquisition for not accommodating a certain kind of generalization (which humans appear to exhibit). Such criticisms might, indeed, be warranted (as in Hadley's case; see chapter 4), but it is most often not acknowledged that symbolic models also suffer from the same (or very similar) deficits. This problem of 'not seeing the splint in one's own eye' is further aggravated by the fact that most connectionists models are *implemented* as computational models, whereas many of their symbolic counterparts are not, but remain *conceptual* models. This means that the empirical results of connectionist simulations are often compared directly with conceptual predictions from an unimplemented symbolic model (as it is, e.g., in the case of Pinker & Prince, 1988, and Prince & Pinker, 1989). Of course, it is possible that the symbolic models when implemented would provide the

[1]For example, much of the criticism of connectionism launched by Fodor & McLaughlin (1990) as well as Fodor & Pylyshyn (1988) does not stem from inconsistencies or incoherence *within* the theoretical framework of connectionism. Instead, it stems from the failure on behalf of Fodor and collaborators to couch connectionism in the terminology of the classical processing paradigm (also cf. van Gelder, 1991). Similarly, another non-classical approach to cognition—situation theory—has also been victim of the same kind of terminologically based criticism: "Fodor thinks that computation is formal. So when I argue that thought is not *formal*, he annoyingly charges me with claiming that thought are not *computational*. I suppose Fodor is so caught up in his own identification of formal with computational as to be unable to maintain the distinction" (Barwise, 1989: p. 156–7).

right empirical results; but in the absence of such implementations, the conceptual models should not be taken as existence proofs of classical solutions (given that they can always be modified in some *ad hoc* fashion to accommodate problematic data).

This leads to the last methodological point that I want to make before outlining the content of my thesis. The presupposition of classical theoretical constructs in cognitive explanation and the mismatched comparisons between connectionist and symbolic models, together, may lead to in principle arguments against connectionist models subserving certain (or all) cognitive functions. Consider, for instance, the following summary from Pinker & Prince (1988):

> we will conclude that the claim that parallel distributed processing networks can eliminate the need for rules and rule induction mechanisms in the explanation of human language is unwarranted. In particular, we argue that the shortcomings are in many cases due to central features of connectionist ideology and irremediable; or if remediable, only by copying tenets of the maligned symbolic theory. The implications for the promise of connectionism in explicating language are, we think, profound. (p. 82)

This is, perhaps, not an 'in principle-argument', but it comes very close. In any event, the problem here is that the empirical shortcomings of *a particular* connectionist model performing some function is generalized to a principled argument against *all* connectionist models of that given function. However, only additional empirical research can establish whether such shortcomings are endemic to connectionist networks, or merely an artifact of a particular implementation. In closing these methodological preliminaries regarding the incommensurability trap, it is worth noting that the opposite danger is equally real—symbolic models can look unattractive from a connectionist perspective. This raises the danger of ignoring all that has been learned from the symbolic approach, and simply starting the project of understanding the mind afresh. In this thesis, I hope to keep clear of this danger.

1.2 Content

One of the dominating characteristics of much connectionist work is the combined emphasis on learning and processing—in particular, when it comes to connectionist models of language. In contrast, much work within the classical approaches to language invoke Chomsky's (e.g., 1980, 1986, 1988, 1993) notion of Universal Grammar (UG); that is, a massive body of innately specified, language specific knowledge. The main challenge for connectionism is thus to account for the variety of linguistic phenomena which traditionally have been taken to support a nativist position on language. Recently, the

general relation between connectionism and nativism has been the focus of some debate (e.g., Christiansen & Chater, 1992; Clark, 1993; Quartz, 1993), especially in connection with language acquisition (e.g., Bates & Elman, 1993; Elman, 1993; Karmiloff-Smith, 1992; Kirsh, 1992; Narayanan, 1992; Ramsey & Stich, 1991). One of the important conclusions to be drawn from this debate is that connectionist models do not have to be, and in most cases are not, *tabula rasa* models of learning. More specifically, it is important to distinguish between domain-specific and more general innate constraints on connectionist learning. In this thesis, I present a connectionist theory of the use, acquisition and evolution of language, incorporating innate, but *not* language-specific constraints on processing and learning. As such, the theory provides a learning and processing based alternative to UG approaches to language.

First, in chapter 2, I address the issue of how much processing power is required to account for human language behavior. A distinction is made between what I call *iterative recursion*, which can be subserved by a finite-state automaton (FSA), and *non-iterative recursion*, for which at least a push-down automaton (PDA) is needed. The existence of the latter kind of recursion has been used as evidence against FSA models of language. However, an unconstrained PDA can process sentences that are beyond the abilities of humans—most notably, in English sentences with multiple center-embeddings. To solve this problem a distinction is typically made between a limited language performance and an infinite linguistic competence. This distinction has generally served as a major obstacle against processing based approaches to language, and against connectionist language models in particular. I therefore challenge the validity of the competence/performance distinction, stressing that it is impossible to distinguish in a theory-neutral way between evidence pertaining to linguistic performance and evidence pertinent to language competence. In addition, the distinction makes the grammar functionally independent of processing, thus threatening to make linguistic theories immune to potentially falsifying empirical evidence.

Having argued against the competence/performance distinction, I turn my attention to the amount of recursion that can be observed in actual language behavior. Psycholinguistic results show that only very limited instances of non-iterative recursion can be found in naturally occurring language. However, a possible objection is that performance on these kind of recursive structures—such as, center-embedded sentences—can be improved given training and external memory aids. I rebut this objection, suggesting via an analogy with the processing of garden path sentences (i.e., structurally ambiguous sentences) that such performance improvements are not rooted in an unbounded linguistic competence, but in the conscious use of reasoning abilities that are not specific to language. The underlying idea is that under normal circumstances we process

language without conscious effort, but when sentences of a certain complex nature are met (such as, garden path and multiply center-embedded sentences), processing fails and higher level reasoning processes are recruited to complete the parse.

From this I conclude that language models in general do not need to account for unlimited non-iterative recursion. Rather, the models only have to be able to deal with a very limited amount of non-iterative recursion, but must also encompass a substantial capacity for the processing of iterative recursive structures. Connectionist models seem to fit this mold. Furthermore, they have no straightforward separation of competence and performance, making them ideal models of language processing vis-a-vis the problems facing more traditional approaches. I therefore discuss in some detail the nature of neural network processing and distributed representation as related to the learning of linguistic structure, paving the way for the simulation experiments reported in chapters 3 and 4.

The existence of non-iterative recursive structure in natural language was one of the principal, and most telling, sources of difficulty for associationist models of linguistic behavior. It has, more recently, become a focus in the debate surrounding the generality of neural network models of language, which many would regard as the natural heirs of the associationist legacy. Non-iterative recursive sentence constructions are difficult to process because it is necessary to keep track of arbitrarily many different dependencies at once. This is not possible for associationist accounts, which assume that the language processor is a (particular kind of) FSA. Similarly, assuming, as we must, that all parameters have finite precision, any finite neural network is also a finite state machine. The important question is, then, whether neural networks can learn to handle non-iterative recursive structures? If not, many would argue, neural networks can be ruled out as viable models of language processing. Chapter 3 therefore investigates to what extent connectionist models are able to learn some degree of non-iterative recursion.

Chomsky's (1956, 1957) original proof against FSA accounts of language was presented in terms of three non-iterative recursive languages. In a number of simulation experiments, I use slightly altered versions of these languages to test the ability of two kinds of recurrent neural networks to acquire non-iterative recursive regularities. Network performance is further measured against a simple statistical prediction method based on n-grams, strings of consecutive words. I outline a number of performance expectations given the complexity of the three test languages, and report the results of two experiments involving, respectively, a 2-word and an 8-word vocabulary. Finally, a comparison is made between network performance and human performance on the same kind of structures, suggesting that both exhibit a similar break-down pattern in performance as the level of embedding increases. I therefore conclude that the existence

of limited non-iterative recursion no longer can be used as *a priori* evidence against connectionist models of linguistic behavior.

In chapter 4, I present simulation experiments which extend the results from the previous chapter. That is, I investigate how a simple recurrent network deals with limited non-iterative recursion in the context of a natural language grammar also incorporating iterative recursion. Results from two experiments are reported: one simulation involving the combination of center-embedded structures with left- and right-branching constructs, and another combining cross-dependency sentence structures with left and right recursive structures. The results show that network behavior on cross-dependency and center-embedded structures in these simulations is comparable with that reported in chapter 3, and therefore with human behavior on similar sentence constructs. Moreover, results pertaining to the model's behavior on the iterative structures is presented, making certain predictions regarding human behavior on such structures.

If connectionist models are to be genuine candidates as mechanisms subserving human language acquisition, it is important that they be able to generalize from past experience to novel stimuli. In this connection, Hadley (1994a) has recently attacked connectionist models of language learning for not affording a sufficiently powerful kind of generalization compared with humans. I discuss this challenge in the second half of chapter 4, and recast it in a more formal and precise way. Then I report additional results from the above simulations, indicating that connectionist models can afford a more sophisticated type of generalization. It is therefore concluded that connectionist models may indeed have sufficient computational power to serve as models for language learning.

Having argued for a processing based view of our language ability and demonstrated the computational adequacy of connectionist models of language, I sketch a theory of the evolution and acquisition of language in chapter 5. The theory distances itself from the dominating exaptationist (e.g., Piattelli-Palmarini, 1989) and adaptationist (e.g., Pinker & Bloom, 1990) perspectives on language evolution by construing language as a *nonobligate symbiant* (that is, a kind of beneficial parasite). As such, there has been a much stronger pressure on language to adapt to the constraints imposed by human learning and processing mechanisms, than *vice versa*. This view promises to provide functional explanations for apparently arbitrary linguistic principles—such as, subjacency—which appear to be universal to all human languages.

I then speculate that language originated as a manual language of gestures, subserved by evolutionary ancient implicit learning processes presumably seated somewhere near Broca's area in the left hemisphere of the brain. Changes in the human

vocal tract are likely to have facilitated the shift from a manual language to a predominately vocal language. The key to understanding subsequent linguistic change, I argue, is vocabulary growth, forcing a gradually more regularized morphology in order to accommodate the growing number of words in a finite memory system. This, in turn, lead to a more complex syntax to ensure reliable communication. Moreover, it is contended that the Baldwin effect, which allows learned traits to become innate, does not apply in the case of language evolution.

Since the evolutionary scenario suggests that learning still plays the dominating role in language acquisition, I close the chapter by outlining a maturationally based theory of language learning. Importantly, language has evolved to be learnable by human infants undergoing maturational changes. It is therefore not surprising that children, despite their limited memory and perceptual abilities, are better language learners than adults. This explains the existence of a critical period of language acquisition. The maturational picture provides the basis for a re-appraisal of the poverty of stimulus argument in all its instantiations; that is, the existence of arbitrary linguistic universals, noisy input, infinite generalization from limited input, the early emergence of many linguistic principles in child language, and the inadequacy of empirical learning methods. After having shown that the primary linguistic stimulus may not be as poor as assumed in theories of UG, I finally respond to possible objections based on, respectively, creolization and the purported existence of a 'morphology gene'.

The conclusion in chapter 6 sums up the main arguments presented in this thesis in favor of processing and learning based theory of the human language ability and its evolutionary past. This theory has summoned evidence from not only connectionist modeling and psycholinguistics, but also anthropology, implicit learning theory, evolutionary theory, and neuroscience. It makes a number of empirical predictions, some of which are presented in the conclusion along with proposals for their further investigation via connectionist simulations or psycholinguistic experiments. Obviously, much still needs to be said and done, but I believe that in this thesis I have taken us a few steps closer to a learning and processing based theory of language.

Chapter 2

Grammars and Language Processing

Since the emergence of generative grammar, language has been construed predominately as a paradigmatic example of the symbolic view of cognition (e.g., Fodor & Pylyshyn, 1988; Pinker & Prince, 1988). However, the conception of cognition as symbol processing arguably grew out of Chomsky's (1959b) attack on Skinner's (1957) behavioristic account of language. Following this attack, the statistical approach to language advocated by the behaviorists quickly became anathema. Instead, theories of language couched in terms of symbols governed by recursive rules became the focus of mainstream linguistics. This cleared the way for a general dismissal of behaviorism and the subsequent dominance of the symbolic paradigm, leading to the birth of cognitive science (and artificial intelligence).

Within the symbolic view of cognition, the apparently unbounded complexity and diversity of natural language, contrasted with the finite character of our cognitive resources, is often taken to warrant a *recursive* language processing mechanism. Nevertheless, it is important to consider whether it is *necessary* to postulate a recursive language mechanism in order to account for language productivity. What seems to be needed in the first place is a mechanism which is able to generate, as well as parse, an infinite number of natural language expressions using only finite means. Obviously, such a mechanism has to be of considerable computational power and, indeed, recursive rules provide a very elegant way of achieving this. Consequently, recursion has been an intrinsic part of most accounts of natural language behavior—perhaps due to the essentially recursive character of most linguistic theories of grammar.[1]

[1]For example, in GB (e.g., Chomsky, 1981) the underlying principles of $\overline{X}$-theory are recursive, as

The existence of recursion in natural language is often taken to be a major stumbling block for connectionist models of linguistic processing—but this is, strictly speaking, not true. Neural networks (and other finite state automata) are able to process a certain kind of recursion which I will refer to as *'iterative recursion* (i-recursion)[2]. This kind of recursion permits the iteration of a given structure either by branching to the left (such as, the multiple prenominal genitives in *'Anita's cat's tail'*) or to the right (as, for instance, the multiple sentential complements in *'I thought Anders said he left'*). The real obstacle for connectionist language processing is therefore not recursion *per se*, but rather the complex forms of recursion that here will be referred to as *'non-iterative recursion'* (ni-recursion). One of the most important kinds of ni-recursion is center-embedding as exemplified in the doubly center-embedded sentence *'the girl that the cat that the dog chased bit saw the boy'*. As such, ni-recursion allows the generation of structures which cannot readily be redescribed in terms of iteration. This implies that no finite state automaton (FSA) can capture ni-recursion—at least, not without additional memory structures. The same also seems to apply to connectionist models. But, does this mean that connectionist networks are not viable as models of language processing? What does the empirical data tell us about the kind of computational power we may need to account for human language behavior?

In this chapter, I address these questions by first discussing the close relation between recursion and linguistic grammars, illustrated by a detailed example involving the parsing of a ni-recursive sentence. An unrestricted parser of this kind can process sentences beyond any human capacity. A distinction is therefore typically made between the finite observable performance of humans and the infinite competence inherent in the parser's recursive grammar. Section 2 challenges this competence/performance distinction on methodological grounds, suggesting that it be abandoned. Next, psycholinguistic results are presented which show that only a very limited amount of ni-recursion occurs in natural language, whereas i-recursion seems to abundant. In this connection, it is argued that performance improvements following training on ni-recursive sentences reflect aspects of higher level cognitive reasoning, rather than unveiling an underlying unbounded competence. Finally, connectionism is proposed as a processing framework which may permit the eschewal of the competence/performance distinction, while promising to have sufficient computational power to deal with both i-recursion and a limited degree of ni-recursion.

---

are the ID-rules of GPSG (Gazdar *et al.*, 1985).

[2]Note that I am not talking about the recursive *languages* of automata theory, but about recursive *structure* in natural languages. Thus, the use of the adjective 'recursive' throughout this thesis refers to the latter structural meaning. My use of 'recursion' will be explicated below.

2.1 Grammars and Recursion

The creative nature of natural language use—its productivity—has been assumed to be subserved by a system of rules for more than 150 years (cf. Chomsky, 1965). Yet, the generative aspect of grammar had to wait for the development of recursion theory in the foundations of mathematics before it could be properly captured in mechanistic terms. In other words, "although it was well understood that linguistic processes are in some sense 'creative', the technical devices for expressing a system of recursive processes were simply not available until much more recently" (Chomsky, 1965: p. 8). Thus recursion has been an intrinsic part of the generative grammar framework from its inception.

The history of generative grammar dates back to Chomsky's (1956, 1957) demonstration that language can in principle be characterized by a set of recursive phrase structure rules, complemented by a set of transformational rules. There are three ways in which recursion can occur in a phrase structure rule. Suppose X is a (non-terminal) symbol and α and β non-empty strings of symbols and terminals[3]. A rule involves *left-embedding* when $X \Rightarrow X\beta$ (i.e., there is a derivation from X to $X\beta$), *center-embedding* when $X \Rightarrow \alpha X\beta$, and *right-embedding* when $X \Rightarrow \alpha X$. The mechanistic realization of such recursive phrase structure grammars imposes certain demands regarding computational power. For example, Chomsky (1956, 1957) has argued that the language processing mechanism cannot be realized by a finite state automaton (FSA). This is because the latter can only produce regular languages and these are not able to support an unbounded depth of center-embedding. In contrast, the set of human languages is generally assumed to fall (mostly) within the broader class of languages, referred to as 'context-free' within the Chomsky (1959a) hierarchy of languages.

This hierarchy of languages is couched in terms of grammars comprising a number of rewrite rules of the general form $\alpha \rightarrow \beta$. Every class within the hierarchy is defined in terms of the restrictions imposed on its rewrite rules. At each level a given class forms a proper subset of the less restrictive class(es) above it. Restricting α to be a single symbol and β either a single symbol followed by a (possibly empty) string of terminals, or *vice versa*, then we obtain the most narrow of the classes: the *regular languages* which can be generated by an FSA. For example, a *right-linear* rewrite rule takes the form $A \rightarrow wB$ (assuming that A and B are symbols, and w a string of terminals, possibly empty), whereas a *left-linear* rule will have the format of $A \rightarrow Bw$[4]. The less restrictive class of *context-free languages*[5] allows β to be any (non-empty) string

[3]Here, *terminals* correspond to words and *symbols* to variables ranging over the left handsides of the set of rewrite rules (including themselves).

[4]Notice that if $A = B$ then we get right and left recursion, respectively.

[5]The classes of languages from context-free and upwards in the hierarchy all support ni-recursion,

S	→	NP VP
NP	→	N rel
NP	→	N
VP	→	V
rel	→	NP V(trans)

Figure 2.1. The recursive set of phrase structure rules used to assign syntactic structure to sentences with object relative clauses.

of symbols and terminals, making center-embedding possible, e.g., $A \rightarrow w_1 A w_2$. A push-down automaton (i.e., an FSA augmented with a stack—see below) is needed to produce the set of non-regular languages in this class. The *context-sensitive languages* constitute an even broader class, loosening the restriction on α and, thus, allowing more than one symbol to occur on the lefthand-side (though still ensuring that β is at least of the same length as α). As an example, consider the rewrite rule $A_1 B_1 \rightarrow A_1 A_2 B_1 B_2$ which takes $A_1 B_1$ and expands it such that the dependency between A_1 and B_1 crosses that of A_2 and B_2. To generate the set of these languages that are not also context-free, we need a linear bounded automaton (that is, a Turing machine whose tape is linearly bounded by the length of the input sentences). Finally, the broadest class of *unrestricted languages* does not have any restrictions on α and β, and can only be produced in its entirety by a Turing machine with an arbitrarily long tape.

2.1.1 A Parsing Example

If Chomsky (1956, 1957, 1959a) is right in that we need the power of context-free languages in order to account for English (and other natural languages), we might ask what kind of computational mechanism is warranted? As mentioned earlier, the natural language processing mechanism has traditionally been construed as a symbolic processing device using *recursion*. As such, recursion entails that the non-terminal symbol (α) on the lefthand side of a rule reappears (in the β-string) on the righthand side of the same or another rule (as described above). In addition, recursion often occurs as a consequence of the application of a set of rules, each of which by themselves is not recursive. For example, assigning syntactic structure to the (ni-recursive) string *'boys girls cats bite see fall'* (and recognizing it as a complex sentence involving two center-embeddings) is readily done using the *recursive set* of phrase structure rules in Figure 2.1. An analysis of the string in terms of these phrase structure rules would

save their subset of regular languages which only permits i-recursion.

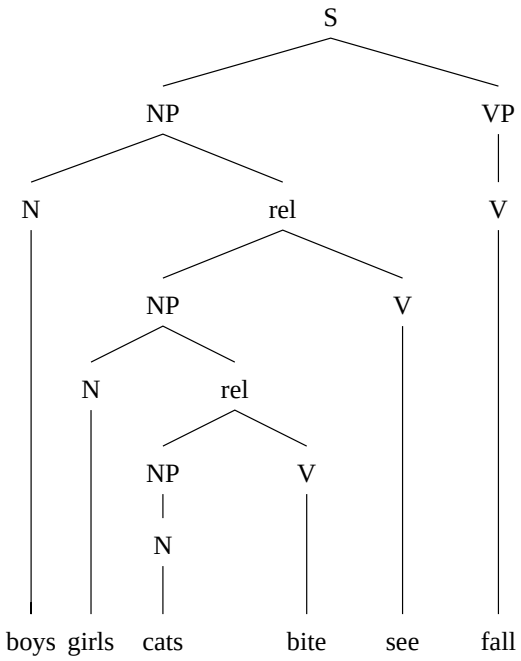


Figure 2.2. The syntactic tree for the doubly center-embedded sentence *'boys girls cats bite see fall'*.

result in the syntactic tree illustrated in Figure 2.2.

At this point one might wonder how we can recover the syntactic information as exemplified by Figure 2.2 from the set of phrase structure rules listed in Figure 2.1. Or, more generally, how can the assignment of syntactic structure given a particular grammar be mechanized so as to allow the processing of arbitrary utterances? What we need is a mechanism—a *parser*—which in some way realizes a grammar such that it is able to recover the syntactic structure of (grammatical) utterances. As such, parsers are typically built as rule-based production systems (Newell & Simon, 1976)[6], comprising a knowledge base realizing the grammar, a working memory in which intermediate processing results are stored, and a mechanism for applying the grammar rules given the content of the working memory. Non-iterative recursion is typically implemented by configuring part of the working memory as a last-in-first-out push-down store, also known as a *stack*. This data structure can be visualized as a pile (or stack) of papers that can only be manipulated in two ways: either, we can remove a single piece of paper from the top of the pile, or, place yet another piece of paper on top. The last

[6]For example, Marcus (1978) acknowledges this when describing his parser, PARSIFAL, in which the grammar "is made up of pattern/action rules; this grammar can be viewed as an augmented form of Newell and Simon's production systems" (p. 237).

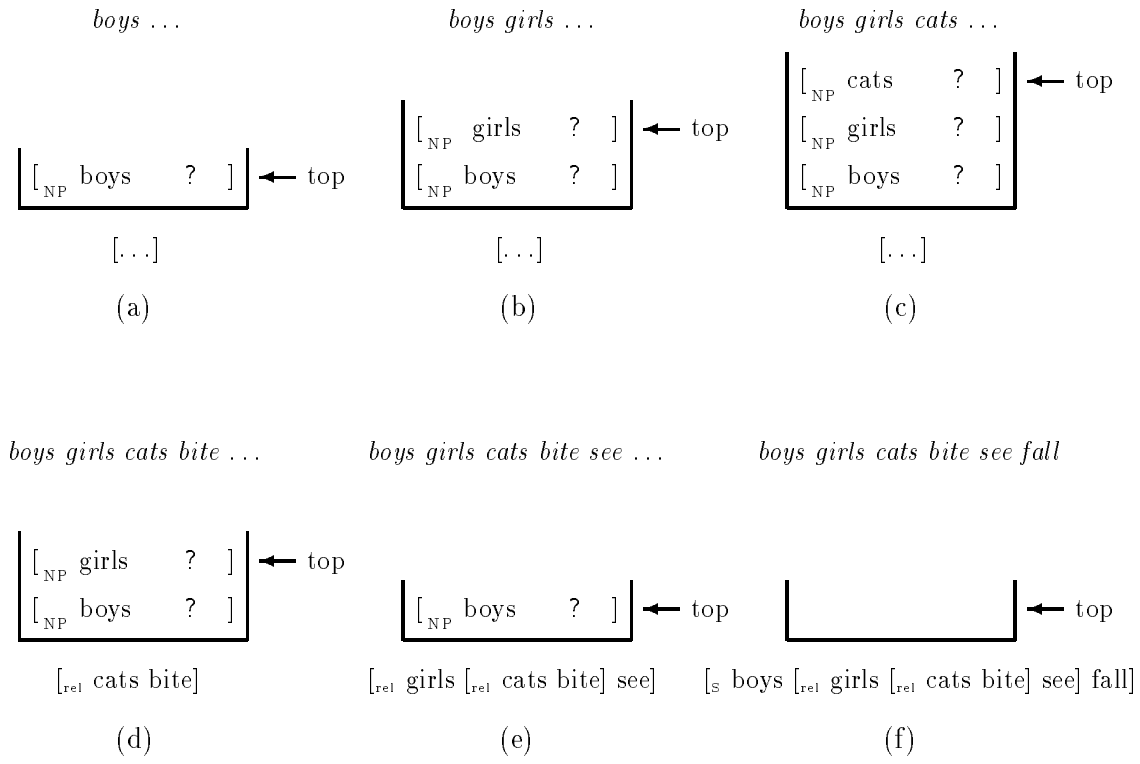


Figure 2.3. Stack (above) and constituent buffer (below) contents when parsing the doubly center-embedded sentence *'boys girls cats bite see fall'*

paper to be placed on top of the pile will therefore be always be the first paper that can be accessed. Notice that just as it is not possible to read papers that are *in* the pile, the parser typically only has access to the element on top of the stack.

To provide a rough idea of how ni-recursion is implemented via a stack, consider the following simplified version of a left-to-right parse of the above sentence. For the purpose of clarification, the parser is not equipped with lookahead buffer. Figure 2.3 shows the stack contents during the processing of *'boys girls cats bite see fall'*. In (a), the parser receives *'boys'* as input, and categorizes it as a noun that can instantiate the N variable in either of the two NP rules from Figure 2.1. Since the parser cannot do anything with just a single noun, a partly instantiated NP rule is 'pushed' on the stack (leaving the constituent buffer empty). Next, the parser gets the noun *'girls'* as illustrated in (b). The parser is unable to attach the two nouns to each other via a rule, so it is forced once more to push a partly instantiated NP rule on top of the stack. When the parser subsequently encounters *'cats'* in (c), it must again push a partly instantiated NP rule on the stack. Following the categorization of *'bite'* as a transitive

verb, the partly instantiated NP rule is 'popped' off the stack (d). The two most recent input words can now form a reduced relative clause, [$_{rel}$ cats bite], in the constituent buffer. This means that rel in the complex NP can be instantiated with *'girls'* as N. The parser then receives the transitive verb *'see'* in (e), and the process is repeated. The constituent buffer now contains [$_{rel}$ girls [$_{rel}$ cats bite] see]. Once again, the next item on the stack can be used to instantiate N in yet another complex NP with the content of the constituent buffer as rel. Finally in (f), the parser gets the last remaining verb, *'fall'*, which matches as a VP with the complex NP to form an S. The constituent buffer now holds the full syntactic parse of the doubly center-embedded sentence (in bracketed form) corresponding to the syntactic tree in Figure 2.2.

From the example, we can see that a stack provides an elegant way of realizing a ni-recursive grammar. Notice that the rules used by the parser above correspond directly to the phrase structure rules in Figure 2.1. However, such a direct identity relation is not required between the parser rules and the rules of the grammar that the parser is realizing. Indeed, this is most often *not* the case for parsers of some complexity (e.g., the parsers developed by Berwick & Weinberg, 1984; Church, 1982; Marcus, 1980). Instead, a general mapping between the theoretical grammar as a whole and its parsing realization is typically adopted (for a more detailed picture, see, for example, Berwick & Weinberg, 1984). Importantly, in this weaker version of the grammar/parser relationship the recursive nature of generative grammar is still mirrored directly in the (recursive) workings of the parser (as exemplified in Figure 2.3—albeit without the one-to-one mapping between grammar and parser rules).

Historically, the direct identity relation was tried out first, ideally providing a basis for a strong link between grammatical competence and parsing performance. This was the motivation behind the *Derivational Theory of Complexity*—first outlined by Miller & Chomsky (1963)—which sought to establish a one-to-one correspondence between grammar rules and the parser's computation over its representations. The basic idea was that the number of transformations required to derive the syntactic structure for a given utterance would be reflected in terms of parsing complexity (that is, the actual time it would take to parse the utterance). This hypothesis was abandoned in favor of more general mappings in the late sixties, following a number of experimental studies showing no direct relationship between the length of tranformational derivation and processing time. However, this move also led to a methodological separation of grammatical knowledge from processing behavior, and subsequently accentuated the competence/performance distinction to which I turn to next.

2.2 The Competence/Performance Distinction

In modern linguistics, the paradigmatic method of obtaining data is through intuitive grammaticality judgments. However, it is a generally accepted fact that the greater the length and complexity of a particular utterance is, the less sure people are in their judgment thereof. Moreover, a variety of psycholinguistic studies have provided much evidence demonstrating the limitations of human language processing (for example, in the case of center-embedded sentences, Bach, Brown & Marslen-Wilson, 1986; Blaubergs & Braine, 1974; Foss & Cairns, 1970; King & Just, 1991; Larkin & Burns, 1977; Marks, 1968; Miller, 1962; Miller & Isard, 1964; Stolz, 1967). In order to salvage the unbounded capacity of the recursively specified generative grammar from the dilemma caused by such human shortcomings, a distinction is typically made between an infinite linguistic *competence* and a limited observable *performance*. The latter is limited by memory limitations, attention span, lack of concentration, and other processing constraints, whereas the former is construed as being essentially infinite in virtue of the recursive nature of grammar and a total lack of constraints on syntactic derivation. Consequently, "a grammar of a language purports to be a description of the ideal speaker-hearer's intrinsic competence" (Chomsky, 1965, p. 4).

The competence/performance distinction has been strongly advocated by Chomsky as having important methodological implications for language research. In particular, he has stressed that it is a common fallacy

> ...to assume that if some experimental result provides counter-evidence to a theory of processing that includes a grammatical theory T and parsing procedure P ...then it is T that is challenged and must be changed. The conclusion is particularly unreasonable in the light of the fact that in general there is independent (so-called "linguistic") evidence in support of T while there is no reason at all to believe that P is true. (Chomsky, 1981: p. 283)

Since this position endorses a sharp functional distinction between linguistic competence and processing performance, I will refer to it as the *strong* C/PD. According to the strong C/PD, linguists do not need to pay much attention to psycholinguistics. Indeed, Chomsky seems to doubt the relevance of psycholinguistic results to language research:

> In the real world of actual research on language, it would be fair to say, I think, that principles based on evidence derived from informant judgment have proved to be deeper and more revealing than those based on evidence derived from experiments on processing and the like, although the future may be different in this regard. (Chomsky, 1980: p. 200)

In this light, the strong C/PD provides its proponents with a protective belt that surrounds their grammatical theories and makes them empirically impenetrable to psycholinguistic counter-evidence.

2.2.1 The Chomskyan Competence Paradox

There is, however, a methodological problem inherent in the strong C/PD. On the one hand, all linguistic theories rely on grammaticality judgments that (indirectly via processing) display our knowledge of language. The strong C/PD, on the other hand, makes T immune to all kinds of empirical falsification—even (in a pinch) to informant judgments—should they not fit T. This leads to what I will call the *Chomskyan competence paradox*. Thus, it seems paradoxical from a methodological perspective to accept only certain kinds of empirical evidence (i.e., grammaticality judgments), whereas the rest is dismissed on what appears to be rather opportunistic grounds. Interestingly, George (1989) has argued against the 'processing fetishism' of cognitive psychology and in defense of the psychological reality of linguistic grammars, that it is a "a peculiar, but common, assumption that we know *a priori* which kind of data (here, 'the behavioral sort') is relevant to which conjectures" (p. 104). Note that this argument (*mutatis mutandis*) applies equally well to the general *rejection* of psycholinguistic data, and, in particular, to the selective use of informant judgments. George's argument is therefore just as dangerous for proponents of the strong C/PD as for cognitive psychologists focusing entirely on behavioral data.

The early history of generative grammar provides an interesting background for the Chomskyan competence paradox. Chomsky originally stated that for an utterance to be grammatical it must be "acceptable to a native speaker" (1957, p. 13). He even suggested a negative test for grammaticality in which an ungrammatical sentence would be read "with just the intonation pattern given any sequence of unrelated words" (p. 16). This ungrammaticality test became problematic in the light of Miller's (1962) results which showed that when subjects are asked to repeat center-embedded sentences, "their intonation is characteristic of the recitation of a list of unrelated phrases, not the utterance of a sentence" (p. 755–6). It was therefore soon abandoned, as was the positive test; i.e., the link between grammaticalness and acceptability. Instead, Chomsky (1965) contended that "acceptability is a concept that belongs to the study of performance, whereas grammaticalness belongs to the study of competence. ...although one might propose various operational tests for acceptability, it is unlikely that a *necessary and sufficient operational criterion* might be invented for the much more abstract and far more important notion of grammaticalness" (p. 11, my emphasis). Having thus severed

the direct link between grammaticalness and acceptability, endorsing the strong C/PD, the opportunistic escape hatch was then further secured: "it is quite apparent that a speaker's reports and viewpoints about his behavior and his competence may be in error. Thus generative grammar attempts to specify what a speaker actually knows, not what he may report about his knowledge" (p. 8). This makes linguistic theory immune towards potentially damaging informant judgments such as those reported in Marks (1968). He found that people judge center-embedded sentences with more than one level of embedding to be ungrammatical (whereas left- and right-embedded sentences of differing length are judged to be grammatical).

The Chomskyan competence paradox should now be obvious: how can it be that amongst informant judgments, which are all elicited in *exactly* the same way, some are considered to be 'true' facts about linguistic competence whereas others merely reflect performance? In other words, how can a theory (of linguistic competence) which builds on a set of facts (i.e., grammaticality judgments) at the same time be used to distinguish among those very same facts, dismissing some of them, without circularity? One could, of course, respond that in certain exceptional cases we can allow (to rephrase Quine, 1960) that the 'tail' of the theory is wagging the 'dog' of evidence. However, this methodological move just leads us to a new version of the problem: how can we determine in a theory-neutral way which linguistic constraints are true limitations on competence and which are merely performance shortcomings? For example, why is it that the subjacency principle (which *can* be explained as a processing constraint cf., e.g., Berwick & Weinberg, 1984; Marcus, 1980) is typically considered to be a genuine constraint on competence, whereas human processing limitations with respect to center-embedded sentences are regarded as performance constraints? This echoed by Church (1982) who—whilst construing subjacency as a competence idealization—nevertheless notes that "in general, though, it is extremely difficult to prove that a particular phenomenon is necessarily a matter of competence. We have no proof that subjacency is a competence universal, and similarly, we have no proof that center-embedding is a processing universal" (p. 12). The Chomskyan competence paradox therefore seems to be an unavoidable methodological implication of the strong C/PD, suggesting that the latter be eschewed[7].

[7]It might also be objected that we need a C/PD to account for the often mentioned fact that people's performance on center-embedded sentences can be improved with training and the use of pen and paper. So, the argument goes, we must have an (at least) in principle infinite competence underlying our linguistic behavior. I will address this argument in section 2.3, suggesting that upon scrutiny it does not hold water.

2.2.2 A Weaker Competence/Performance Distinction

A more moderate position, which I will refer to as the *weak* C/PD, contends that although linguistic competence is supposed to be infinite, the underlying grammar must directly support an empirically appropriate performance. This is done by explicitly allowing performance—or processing—considerations to constrain the grammar. Pickering & Chater (1992) have suggested that such constraints must be built into the representations underlying the grammatical theory, forcing a closer relation to the processing theory. This ensures that the relation between the theory of grammatical competence (Chomsky's T) and the processing assumptions (Chomsky's P) is no longer arbitrary, resulting in an opening for empirical testing. Nevertheless, inasmuch as T and P are still functionally independent of each other, the option is always open for referring any falsifying empirical data questioning T to problems regarding the independent P, i.e., to performance errors. In addition, although this position does not lead directly to the Chomskyan competence paradox, it still relies on a methodologically problematic, theory-laden notion of what counts as evidence of competence and what counts as performance data.

To compare the methodological differences between models of natural language processing that adopt, respectively, the strong or the weak C/PD, it is illustrative to construe the models as rule-based production systems. Recall that within the production system conceptualization of a parser, the grammar of a particular linguistic theory corresponds to the system's knowledge base. The system can therefore be said to have an infinite linguistic competence in virtue of its independent knowledge base, whereas its performance is constrained by working memory limitations. This is in direct correspondence with the strong C/PD, since the grammar is completely separated from processing. Models adhering to the weak C/PD would similarly have an independent, declarative knowledge base corresponding to the grammar, but in addition they would also encompass a separate knowledge base consisting of what we might coin *linguistic meta-knowledge*. This knowledge consists of various performance motivated parsing heuristics—such as, for example, the *'minimal attachment principle'* (Frazier & Fodor, 1978), *'early closure'* (Kimball, 1973), and *'late closure'* (Frazier & Fodor, 1978)—that provide constraints on the *application* of the grammar rules. Thus, the performance of the model is constrained not only by limitations on working memory but also by linguistic meta-knowledge.

From the production system analogy it can be seen that proponents of both the strong and the weak C/PD stipulate grammars that are functionally independent from processing. As a consequence, empirical evidence that appears to falsify a particular

grammar can always be rejected as a result of processing constraints—either construed as limitations on working memory (strong C/PD) or as a combination of working memory limitations and false linguistic meta-knowledge (weak C/PD). In short, as long as the C/PD—weak or strong—is upheld, potentially falsifying evidence can always be explained away by referring to performance errors. This is methodologically unsound insofar as linguists want to claim that their grammars are part of the psychological reality of natural language. It is clear that Chomsky (1986) finds that linguistic grammars are psychologically real when he says that the standpoint of generative grammar "is that of individual psychology" (p. 3). Nevertheless, by evoking the distinction between grammatical competence and observable natural language behavior, thus disallowing negative empirical testing, linguists cannot hope to find other than speculative (or what Chomsky calls 'independent linguistic') support for their theories. In other words, if linguistic theory is to warrant psychological claims, then the C/PD must be abandoned[8].

Having argued against the C/PD, both weak and strong, on methodological grounds, I will now turn to the question of whether the apparent occurrence of different kinds of recursion in natural language after all does warrant talk about a linguistic competence functionally independent from observable performance.

2.3 Recursion and Natural Language Behavior

The history of the relationship between generative grammar and language behavior dates back to Chomsky's (1957) demonstration that language can, in principle, be characterized by a set of recursive rules. Recall that he argued that natural language cannot be accounted for by a finite state automaton, because the latter can only produce regular languages. This class of languages—although able to capture *left-* and *right-*branching i-recursive structures—cannot represent *center-*embedded expressions. For linguistic theories adhering to the C/PD (weak or strong), this restriction on the power of the finite-state grammars prevents them from being accepted as characterizations of the idealized linguistic competence. On this view, natural language must be at least context-free, if not weakly context-sensitive (cf. Horrocks, 1987). However, having eschewed the C/PD on methodological grounds in the previous section, the question still remains concerning how much processing power is needed in order to account for

[8]By this I do not mean that the present linguistic theories are without explanatory value. On the contrary, I am perfectly happy to accept that these theories might warrant certain *indirect* claims with respect to the language mechanism, insofar as they provide means for describing empirical natural language behavior.

observable language behavior. Do we need to postulate a language mechanism with the full computational power of a ni-recursive context-free grammar?

2.3.1 Non-iterative Recursion

Before answering this question, it is worth having a look at some examples of different kinds of recursive natural language expressions. Since the crucial distinction between regular and other richer languages is that the former cannot produce expressions involving unbounded center-embedding, we will look at such ni-recursive sentences first. As the following three examples show, the difficulty of processing a center-embedded sentence increases with the depth of embedding:

(1) The cat that the dog chased bit the girl.
(2) The girl that the cat that the dog chased bit saw the boy.
(3) The boy that the girl that the cat that the dog chased bit saw fell.

The processing of center-embedded sentences has been studied extensively. These studies have shown, for example, that English sentences with more than one center-embedding (e.g., (2) and (3) above) are read with the same intonation as a list of random words (Miller, 1962), cannot be easily memorized (Foss & Cairns, 1970; Miller & Isard, 1964), and are judged to be ungrammatical (Marks, 1968). Bach, Brown & Marslen-Wilson (1986) found the same behavioral pattern in German, reporting a marked deterioration of the understanding and a sharp increase in the negative judgments of comprehensibility of center-embedded sentences with an embedding depth of more than one. It has also been shown that semantic bias and training can improve performance on such structures (Blaubergs & Braine, 1974; Stolz, 1967), and that processing is influenced by individual differences in memory span. Importantly, Larkin & Burns (1977) have furthermore demonstrated that the difficulty in the processing of ni-recursion is not confined to a linguistic context. These findings, in turn, have lead to much debate concerning how they should be incorporated into accounts of natural language processing (e.g., Berwick & Weinberg, 1984; Church, 1982; Frazier & Fodor, 1978; Kimball, 1973; Pulman, 1986; Reich, 1969; Wanner, 1980).

Proponents of the C/PD have explained the difficulty in terms of performance limitations. For example, in order to account for the problems of parsing center-embedded sentences, both Kimball's (1973) parser and Frazier & Fodor's (1978) 'Sausage Machine' parser apply a performance-justified notion of a viewing 'window' (or look-ahead). The window, which signifies memory span, has a length of about six words and is shifted continuously through a sentence. Problems with center-embedded sentences are due to the parser not being able to attach syntactic structure to the sentences because the verb

belonging to the first NP is outside the scope of the window. However, this solution is problematic in itself (cf. Wanner, 1980) since doubly center-embedded sentences with only six words do exist and are just as difficult to understand as longer sentences of similar kind; e.g.:

(4) Boys girls cats bite see fall.

Others (e.g., Church, 1982; Pulman, 1986) impose limitations on the number of items allowed at any one time on the stack and invoke some kind of "forgetting" procedure to ensure the stack does not grow beyond its limit. However, this stack limit is determined in a rather *ad hoc* fashion so as to tailor the behavior of the parser to be comparable with human performance[9].

A possible way out of this problem due to Reich (1969) is to argue that center-embedded sentences, such as (2)–(4), are ungrammatical. However, this suggestion has met strong opposition—even from many psycholinguists emphasizing performance models of natural language (e.g., Berwick & Weinberg, 1984; Church, 1982; Pulman, 1986). The standard counter-argument consists in pointing out that performance can be improved via training and by allowing the use of external memory aids (such as pen and paper) and extra time, whereas this does not appear to be the case for ungrammatical strings. This is therefore taken to reflect a genuine grammatical competence underlying the improvements in performance. Because this argument typically is considered to provide a solid defense of the C/PD, I will address it in detail.

My main objection to the practice/extra resource argument is that it may not tell us anything about the language capacity *per se*. Notice that these performance improvements can only be obtained through hard *conscious* work (e.g., through verbal rehearsing, writing, etc.). In contrast, the processing of natural language under "normal" circumstances is effortless and unconscious. This suggests a process of consciously 'augmenting' an already existing—but relatively limited—grammatical ability[10], rather than unveiling parts of an underlying infinite competence.

A Lesson from Garden Path Sentences

This objection is inspired by Marcus' (1980) *'determinism hypothesis*:

[9]It should be noted that even though Pulman's parser uses an unconventional 'stack-like' memory structure, in which items *within* the stack can be accessed, the limitation on the number of items on this stack is still determined in an *ad hoc* fashion.

[10]This process of conscious augmentation may rely on abstract knowledge of language acquired through schooling and/or semantic and contextual information about which lexical items may go together. The latter has also been proposed by Stolz (1967), suggesting that subjects "might 'learn' the [center-embedded] construction by noticing some connection between the test sentences and aspects of his nonlinguistic knowledge during the course of the experiment" (p. 869).

> There is enough information in the structure of natural language in general, and in English in particular, to allow left-to-right deterministic parsing of those sentences which a native speaker can analyze *without conscious effort.* (Marcus, 1980, p. 204; my emphasis)

In this context, 'determinism' means that the language processor is committed to whatever syntactic structure it is presently building, thus disallowing deterministic simulations of non-determinism such as backtracking and pseudo-parallelism (i.e., the possibility of following several different syntactic paths simultaneously)[11]. Marcus proposed the above hypothesis in order to be able to predict which sentences would send people "down the garden path". Garden path sentences are fully grammatical sentences in which a structural ambiguity encountered early in the sentence might lead the reader/listener to go down the wrong syntactic path. Consider, for example, the classic garden path sentence:

(5) The horse raced past the barn fell.

In (5), the word *'raced'* induces a potential structural ambiguity, since it can be parsed either as the main verb of a past tense sentence, or as the start of an unmarked reduced relative clause (i.e, as a past tense participle). The former parse leads the language processor down the garden path, because it will not be able to fit *'fell'* into the syntactic structure that it is building. Interestingly, most people will parse *'raced'* in this way, as if they were expecting the sentence to look like (6) rather than (5)[12].

(6) The horse raced past the barn.

So, when they encounter *'fell'* they become aware of their misparsing, backtrack, and then try to reparse the sentence from scratch[13].

---

[11] It may be that this strict determinism will have to be replaced by a more qualified determinism as found in the recent multiple-constraint models of parsing and comprehension (e.g., Taraban & McClelland, 1990). Such a change would, however, not change the general gist of the present conscious augmentation hypothesis.

[12] The preference of the main verb reading over the participle reading of *'raced'* might be explained in terms of the distributional statistics concerning the relative occurrence of the two verb forms in everyday language. We may therefore expect that the main verb (past tense) reading occurs significantly more often than the participle reading, leading to a strong bias towards the former in ambiguous cases such as (5). Thus, (6) is more likely to occur in normal English than (5). Evidence presented in, e.g., Juliano & Tanenhaus (1993) corroborates this expectation

[13] Note that the garden path in (5) can be avoided if the relative clause is marked and unreduced as in: *'The horse that was raced past the barn fell'* (making it a *center-embedded* sentence of depth one). Moreover, experiments have shown that semantic (Milne, 1982) and contextual information (Altmann & Steedman, 1988; Crain & Steedman, 1985; Taraban & McClelland, 1990) can both induce and prevent garden paths.

The observable behavior of people, when faced with garden path sentences, have lead Marcus (1980) to propose his determinism hypothesis, and subsequently Milne (1982) to suggest that

> ...the processing of a normal sentence does not require conscious effort and [it] is generally agreed that to understand a garden path sentence requires conscious effort. The reader notices a mental 'jump' or 'block' when reading of the sentence, stops and the garden path is consciously realized. Experimentally, conscious effort is detected by an increase in reaction time to a given task. As an armchair definition; any grammatical sentence that seems abnormal to read, requires conscious effort. (p. 353)

The operational criterion for 'conscious effort' has been confirmed by response time studies (Milne, 1982) demonstrating significantly longer reading times for garden path sentences, such as (5), compared with the corresponding non-garden path sentences, such as (6) (the latter being slightly altered to control for sentence length: *'The horse raced past the old barn'*). Eye-tracking studies further support the idea that conscious effort may be needed to recover from the garden path. For example, Rayner, Carlson & Frazier (1983) found that people make regressive eye movements when experiencing the garden path effect in (5), indicating that conscious backtracking may be necessary to recover from certain garden path sentences. However, there are also garden path sentences which produce longer reading times, but which nevertheless do not seem to require conscious effort[14]. This may be paralleled by the processing of sentences with a single center-embedded object relative clause, such as (1), which do not appear to elicit conscious awarenes. These sentences also produce longer reading times in comparison with sentences expressing the same sentential content using a subject relative clause (cf. King & Just, 1991), as it was the case with the 'unconscious' garden path effect.

The lesson to be learned from the garden path case is that syntactic processing (in Milne's words) "is unconscious, deterministic and fast, but limited" (1982, p. 372). Despite the limitations, syntactic processing rarely fails. But when it does, conscious effort is needed to recover from the error. This allows the allocation of non-language specific, cognitive resources so that conscious re-processing of the problematic utterance is made possible.

Center-embedding and Conscious Processing

Returning to the case of center-embedded sentences, I submit that when people are presented with these kind of language constructions they exhibit the same kind of

[14]In this connection, it should be noted that conscious awareness is presumably not an all or non phenomena.

behavior as with garden path sentences. For example, Miller & Isard (1964) report work by Mackworth & Bruner in which the latter "have recorded the eye movements of subjects when reading these sentences and have found that the number of fixations and the number of regressive movements of the eyes both increase markedly as the amount of self-embedding increases" (p. 299). Miller & Isard also go on to note the direct correlation between conscious awareness and the point in the sentence where the grammatical complexity starts:

> Mackworth and Bruner's recordings of eye movements while reading such sentences confirms the *introspective impression* that the difficulty does not begin until the long string of apparently unrelated verbs is encountered toward the end of the self-embedded sentence. At this point the recursive eye movements begin and *one feels* that all grasp of the sentence has suddenly crumbled away" (p. 301; my emphasis).

It should be clear that these behavioral findings are similar to evidence concerning the processing of garden path sentences (recall the "mental jump" mentioned by Milne, 1982, and the regressive eye movements reported by Frazier, Carlson & Rayner, 1983). Thus the processing of center-embedded sentences is unconscious and effortless until the second or third verb is encountered towards the end of the sentence. At this point conscious effort becomes necessary so that the string of verbs can be combined with the correct NPs in the beginning of the sentence. Notice that the first verb (e.g., *'chased'* in (1)–(3)) can easily be combined with the last NP (i.e., *'the dog'* in (1)–(3)). The results of Bach, Brown & Marslen-Wilson (1986), Marks (1968), Miller (1962), and Miller & Isard (1964) taken together suggest that the combination of the second verb and the last but one NP should not provide too much difficulty either (so that (1) would still be processed without the need of additional conscious processing). But as soon as a sentence has more than one embedding (i.e., (2)–(4)), thus requiring the combination of more than two NPs with their respective VPs, then the language processor will not be able to complete the parse without additional conscious processing.

But what about the performance improvements on center-embedded sentences obtained through training and/or the addition of extra processing resources? Such improvements are in contrast with the often made observation that "no amount of practice or artificial aids can make a clearly ungrammatical sentence like 'on mat cat the sat the' become grammatical" (Pulman, 1986, p. 204). Although this is trivially true, it is begging the question at hand. The real question, when making a comparison with the center-embedding case, must be whether people would also be able to improve their performance on ungrammatical sentences (and not whether the latter can become grammatical). When addressing this question it is important to keep in mind

that even *with training* the human performance on center-embedded structures is very limited (Blaubergs & Braine, 1974). Indeed, experimental data seem to suggest that many "listeners can decode a novel [ni-]recursive sentence *only* when its recursive quality has been learned as a specific part of the specific grammatical rule involved" (Stolz, 1967: p. 872; my comment and emphasis).

It is often noted that Miller & Isard have demonstrated that free recall of center-embedded sentences can be improved via learning, but it is almost equally often overlooked that they also reported a positive learning effect on strings of random words. In fact, the increase in performance through learning was three times higher for the ungrammatical strings (about 100% from trial 1 to 5) than for each of the center-embedded sentences (with 0–4 embeddings). Since learning can improve performance on both grammatical center-embedded sentences and ungrammatical strings of random words, the former cannot be assigned a special status as evidence in favor of an unbounded competence[15]. That is, the learning induced performance improvements on center-embedded constructions do not support a distinction between competence and performance. Presumably, we do not want to claim that the improvement of performance on the ungrammatical strings is evidence of the language processor's ability to learn ungrammatical constructions (which would be a direct consequence of maintaining the traditional practice/extra resource argument). Instead, we can explain both kinds of performance improvement as pertaining to the conscious use of non-language specific cognitive resources. When the language processor is facing either center-embedded sentences (with more than one embedding) or strings of random words, its unconscious processing is interrupted and the help of our general cognitive reasoning abilities is recruited to complete the task.

This picture would be threatened if we could find naturally occurring sentences with a depth of more than one. De Roeck, Johnson, King, Rosner, Sampson & Varile (1982) claim to have found such evidence. They provide examples from both German and English of sentences with up to six levels of embedding. Importantly, all these examples are from *written texts* (and, furthermore, mostly from sources (in)famous for their complicated constructions). These examples are most certainly conscious

[15]In this connection, I predict that people might also be able to improve their comprehension of ungrammatical sentences (such as, *'on mat cat the sat the'* and more complex cases) via learning. The only restriction necessary on a set of ungrammatical sentences would be that they must all reflect the same ungrammatical regularities (such as, for example, similar incorrect word order). The work on the learning of artificial grammars (for an overview, see Reber, 1989) can be taken to support this prediction, since the former essentially involves the learning of a particular set of regularities (as it would also be the case for the latter). Furthermore, in chapter 5 I argue that such learning processes may subserve our language ability

products of their authors, and also seem to require considerable conscious effort on behalf of the reader to be understood[16]. Such sentences would presumably lead to the characteristic regressive eye movements as mentioned above, suggesting an interruption of the language processor, and the subsequent backtracking and conscious reprocessing.

De Roeck *et al* (1982, p. 332) do however provide one spoken example of a center-embedded sentence of depth two:

(7) Isn't it true that example sentences that people that you know produce are more likely to be accepted?

This sentence appears to be parsable without any problems, thus potentially causing a problem for the above picture of the processing of center-embedded sentences. I contend, however, that this is not the case, because the relative clause *'people that you know'* arguably is parsed as something like a single unit (because of its frequent occurrence in everyday English). This seems to be in accordance with Miller & Isard's (1964) speculation that "subjects could, to a certain extent, organize the discontinuous constituents as single units" (p. 300). In this light, compare (7) with (8).

(8) Isn't it true that example sentences that people that you threaten produce are more likely to be accepted?

In (8), *'people that you threaten'* cannot be as easily chunked into a single unit as it was the case with the innermost relative clause in (7). The subsequent increase in the processing difficulty makes (8) more comparable with (2) than (7)[17]. This difference in processing difficulty between (7) and (8) suggest that semantic, contextual or (as here) distributional statistics might ease the processing of doubly center-embedded sentences—similarly to the case of garden path sentences.

2.3.2 Iterative Recursion

Given the discussion above (and in the previous section), I contend that a language processor need not be able to account for unbounded center-embedding (at any level of analysis). Rather, it should experience difficulty comparable with the experimental

[16] The latter is, e.g., indicated by De Roeck *et al* with respect to their German examples: "The sentences containing multiple center-embeddings ... are, certainly, somewhat clumsy; a German-Swiss colleague commented on ... [one of these sentences], for instance, that it is the kind of sentence which you have to look at twice to understand" (1982, p. 335).

[17] It could be objected that the use of *'threaten'* in (8) makes the sentence semantically odd. However, this oddness does not appear until after all the three verbs have been encountered. Thus, talk about example sentences uttered by people that you threaten does not appear to be less semantically coherent than talk about example sentences uttered by people that you know.

evidence when encountering center-embedded sentences. Still, this leaves left- and right-recursion to be dealt with. That these i-recursive structures cannot be easily dismissed, but seem to be relatively ubiquitous in natural language, can be seen from the following examples involving such constructions as multiple prenominal genitives (9), right-embedded relative clauses (10), multiple embeddings of sentential complements (11), and PP modifications of NPs (12):

(9) [[[[Bob's] uncle's] mother's] cat]...
(10) [This is [the cat that ate [the mouse that bit [the dog that barked]]]].
(11) [Bob thought [that he heard [that Carl said [that Ira was sick]]]].
(12) ...the house [on [[the hill [with the trees]][at [the lake [with the ducks]]]]].

Furthermore, *prima facie* there seems to be no immediate limits to the length of such sentences (but see section 4.2 in chapter 4 for a discussion of possible limitations on i-recursive structures).

Even though (9)–(12) are describable in terms of left- or right-recursion, it has been argued—with support from, e.g., intonational evidence (Reich, 1969)—that these expressions should be construed not as recursive but as *iterative* (Pulman, 1986). Strong support for this claim comes from Ejerhed (1982) who demonstrated that it is possible for an FSA, comprising a non-recursive context-free grammar, to capture the empirical data from Swedish (provided that unbounded dependencies are dealt with semantically). This demonstration is significant because Swedish is normally assumed to require the power of context-sensitive languages (e.g., cf. Horrocks, 1987). Thus, we have strong reasons for believing that an FSA may provide sufficient computational power to account for natural language performance without needing to postulate a functionally independent infinite competence.

In section 2.2, we saw that the competence/performance distinction was a direct consequence of construing the grammar underlying linguistic behavior as a set of recursive rules. Thus generative grammar by its very nature necessarily requires that performance be separated from competence in order to account for observable language behavior. So, it seems that if we are to abandon the distinction between competence and performance (as the previous two sections suggest that we should), then we need a different way of representing our knowledge of language. That is, we need a representational vehicle that will allow us to avoid the methodological problems of the C/PD as well as model syntactic processing as being deterministic and unconscious in compliance with the limitations set by the experimental evidence. In the remaining part of this chapter, I present connectionism as a possible alternative framework in which to represent the regularities subserving our language skills. First I provide a brief account

of the main properties of connectionist representation and processing, before turning to a general discussion of neural networks as models of natural language.

2.4 Connectionist Natural Language Processing

Connectionism typically implies using artificial neural networks as models of psychological phenomena[18]. As such, these networks consist of a collection of simple processing units, typically organized into layers, that are connected to each other by weighted links (for an introduction to connectionism, see, e.g., Bechtel & Abrahamsen, 1991; Clark, 1989; Rumelhart *et al.* , 1986). An important property of neural networks is that they can be trained to develop the representations necessary to deal with a specific task. Moreover, it has been observed that, through learning, connectionist models (with hidden units) are able to develop distributed representations whose internal structure mirrors that of the externally given input. More specifically, vectors corresponding to the individual patterns of activation over the hidden units can be conceived as *points* in a multidimensional state space (e.g., van Gelder, 1991a). The exact location of a given vector is determined by the specific values of its constituents; i.e., by its internal configuration. As a result, similar vectors are mapped into similar locations in space. The degree of similarity between vectors—the 'distance' between them in space—can be measured using a variety of standard vector comparison methods (e.g., cluster analysis or trajectory analysis). Due to the superpositional and highly distributed nature of the networks in question, representations that are structurally similar—i.e., that have similar internal structure or, more precisely, have similar vector configurations—end up as *'neighboring'* positions in state space. Thus, structurally related input representations will invoke 'adjacent' representations in hidden unit state space.

It is important to notice from a computational perspective that these similarities have *causal significance*. The behavior of a network, being a complex dynamical system, is causally dependent on the current pattern of activation over the hidden units; that is, on the current representation's particular location in space. In other words, the specific location in space of a given representation will causally effect how it is processed. Since the internal structure of such distributed representations corresponds systematically and in an essentially non-arbitrary way to the structural configuration of the input representations, allowing us to project any semantic interpretation we might assign the input onto the appropriate positions in vector space, and since variations of position in state space are causally efficacious, the processing of a network can be seen as

[18]The following is based in part on Christiansen & Chater (1992).

being determined systematically according to the semantic content of the distributed representations.

Judging from this exposition it would seem to be the case that connectionist representations can be assigned content in an essentially non-arbitrary way, since their internal structure (given successful training) will correlate with structural contingencies in the input and produce a non-arbitrary representation; that is, connectionist representations appear to be able to possess at least some *bona fide* intrinsic content. However, the internal states of present day connectionist networks appear to be no more "grounded" than their symbolic counterparts (also cf. Bechtel, 1989; Cliff, 1990; Sharkey, 1991). Crucially, the distributed representations in question are only non-arbitrary in relation to the structure of the given input representations, not in relation to what the latter are representations *of*; i.e., the entities they refer to in the outside world. Consequently, similarity is defined as a relation *between* input representations, and not as a relation to the appropriate external objects they are to represent. Furthermore, since the input representations provided by the programmer are typically pre-structured and of a highly abstract nature, it is always possible to give a network's input representations a different interpretation, thus changing the projected content of the internal distributed representations. This has been mirrored empirically by the fact that only a few experiments have been carried out with "real" sensory-type data (in the sense of not having been pre-processed by the programmer), and then with a mostly unsuccessful outcome (see Christiansen & Chater, 1992, for an example and further discussion of this issue).

There is, however, a sense in which connectionist representations are non-arbitrary; that is, the *inter-representational* relations in a network are essentially non-arbitrary. In contrast to symbolic systems in which the atomic symbols have no relation to each other (albeit, complex symbols have non-arbitrary inter-relations), distributed representations are non-arbitrarily related to each other in state space. Whereas atomic symbols designating similar objects have no (non-coincidental) relation to each other, connectionist representations of similar object representations in the input will end up as neighboring points in state space. Thus, connectionist networks provides us with a kind of non-arbitrary representational "shape" that allows a notion of inter-representational similarity. The important ability of connectionist networks to *generalize* derives from these similarity relations between representations corresponding to structurally similar input. Despite the non-arbitrariness of these inter-relations and their grounding of a robust notion of representational similarity, the *extra*-representational links are still fundamentally arbitrary and therefore ungrounded.

2.4.1 Compositionality in Connectionist Models

Turning to the issue of learning linguistic structure, the problem of learning structured representations comes into focus. This problem has received much attention following the debate initiated by Fodor's & Pylyshyn's (1988) attack on connectionism (for example, Chalmers, 1990b; Chater & Oaksford 1990a; Fodor & McLaughlin, 1990; Oaksford, Chater & Stenning, 1990; Smolensky 1987, 1988; van Gelder, 1990a, 1991a). In this connection, it has been suggested that the classical notion of compositionality may be unnecessarily restrictive from the point of view of connectionist systems (that is, the classical understanding of compositionality may induce an instance of the incommensurability trap—forcing connectionist systems into an inappropriate framework). This classical notion is labeled as *concatenative* (or 'syntactic') compositionality, which "must preserve tokens of an expression's constituents (and the sequential relations among tokens) in the expression itself" (van Gelder 1990a: p. 360).

A broader notion, *functional* compositionality, does not demand the preservation of constituents in compound expressions. What *is* needed is a general and reliable mechanism that can produce composite expressions from arbitrary constituents and later decompose them back into their original constituents. As an example of functional compositionality, van Gelder (1990a) points to Gödel numbering, which is a one-to-one correspondence between logical formulae and the natural numbers. For instance, on a given scheme the proposition P will be assigned the Gödel number *32*, whereas a logical expression involving P as a constituent, say, *(P&Q)* would be assigned the Gödel number *51342984000*. It is clear that the Gödel number for *(P&Q)* does not directly (or syntactically) contain the Gödel number for P. Still, by applying the prime decomposition theorem we can easily determine the Gödel numbers for its primitive constituents. Thus, we have constituency relations without concatenative compositionality. Since distributed networks using superimposed representation appear to 'destroys' the constituents of composite input tokens (at least from the viewpoint of human observers), they do not qualify as having concatenative compositionality. However, this is not irreversible because the original constituents can be recreated in the output.

There is a danger that this would leave connectionist representations with the same status as, for example, data-compressed, or otherwise encrypted, files on a standard computer—as being useful only as storage but not for processing. For a genuinely connectionist account of the representation and processing of structured representations, it is necessary to be able to manipulate the functionally compositional representations *directly* as van Gelder stresses. In the case of Gödel numbering, operations which are sensitive to compositional structure (e.g., inferences) will not correspond to a (readily

specifiable) function at the arithmetic level. Hence, performing logical inference over Gödel numbers is a rather hopeless endeavor. Notice too, the compositional *semantics* which can be easily defined over logical representations will have no (readily specifiable) analog at the level of Gödel numbers.

What is important, from the viewpoint of language acquisition and processing, is whether or not connectionist networks can handle (and, in particular, learn to handle) problems which are standardly viewed as requiring structured representations. That is, can connectionist representations attain what we shall call *'apparent'* compositionality. If apparent compositionality can be learned, then there are two possibilities concerning the nature of the representations that the network employs. It could be that, on close analysis, the net is found to have devised a standard, concatenative compositional representation. Alternatively, the network might behave *as if* it used structured representations, without using structured representations at all. In the former case, it would seem appropriate to say that the network representations are compositional (in the standard sense); in the latter, that the network is not using a compositional representation (also in the standard sense). What is required, it appears, is not a new notion of compositionality, but the attempt to devise networks which can behave as if they had structured representations, followed by an analysis of their workings. Of course, there is a third possibility: that representations within networks do, implement compositionality, but in some heretofore unknown way, unlike that used by classical systems (with appropriate operations over it, and an appropriate semantics). *This* possibility would cause us to revise the notion of compositionality, much as the discovery of non-Euclidean geometry enlarged and changed the notion of straight lines, parallel and so on. It will only be possible to develop a specifically connectionist notion of compositionality, or even know if this possibility is coherent at all, *post hoc*—that is, by analyzing networks that exhibit apparent compositionality.[19] In other words, what kind of compositionality we should ascribe connectionist representations is an empirical question, which can only be answered by empirical investigation.

Recently, research efforts have therefore been made towards defining operations that work directly on the encoded distributed representations themselves, instead of their decomposed constituents. Chalmers (1990a) devised a method by which a simple feed-forward, back-propagation network—dubbed a *transformation* network (TN)—was able to manipulate compact distributed representations of active and passive sentences according to their syntactical structure. First, a Recursive Auto-Associative Memory

[19]Of course, it is likely that any such notion would be included as a subclass of functional compositionality (as it is the case with concatenative compositionality)—but functional compositionality *per se* does not put us any further forward to finding such a notion.

(RAAM) network (Pollack, 1988) was trained to encode distributed representations of the sentence structures. Chalmers then trained the TN to transform compact representations of active sentences into compact representations of passive sentences; that is, he trained the network to associate the RAAM-encoded distributed representations of active sentences with their distributed passive counterpart. In similar vein, Niklasson & Sharkey (1992) successfully applied the same combination of RAAM[20] and TN to (a subpart of) the domain of logical axioms. These empirical investigations have shown that it is possible to devise models, such as the TN, that can manipulate the compact distributed representations in a structure sensitive way.

It is worth mentioning that when addressing the issue of connectionist compositionality there is a potential danger of falling into the incommensurability trap. As pointed out by Sharkey (1991), the division between semantics and structural considerations might be somewhat artificial, since such a division seems to be collapsed in much connectionist research. On this view, even our own notion of apparent compositionality could get us trapped in the claws of incommensurability. Nevertheless, bearing this in mind, re-interpretation of old terminology seems to be the only productive way forward for a research program still in its infancy.

Importantly, the above discussion of the general properties of connectionist models suggests that a connectionist perspective on language promises to eschew the C/PD, since it is not possible to isolate a network's representations from its processing. The relation between the 'grammar'—or, rather, the grammatical regularities which has been acquired through training—and the processing is as direct as it can be (van Gelder, 1990b). Instead of being a set of *passive* representations of declarative rules waiting to be manipulated by a central executive, a connectionist grammar is distributed over the network's memory as an *ability to process language* (Port & van Gelder, 1991). In this connection, it is important to notice that although networks are generally 'tailored' to fit the linguistic data, this does not simply imply that a network's failure to fit the data is passed onto the processing mechanism alone. Rather, when you tweak a network to fit a particular set of linguistic data, you are not only changing how it will *process* the data, but also what it will be able to *learn*. That is, any architectural modifications will lead to a change in the overall constraints on a network, forcing it to adapt differently to the contingencies inherent in the data and, consequently, to the acquisition of a different set of grammatical regularities. Thus, since the representation of the grammar is an inseparable and *active* part of a network's processing, it is impossible to separate a

[20] Actually, they applied a slightly modified version of the RAAM in which an additional bit denoted whether the input/output representations were atomic (i.e., not distributed) or complex (i.e., distributed).

connectionist model's competence from its performance.

2.4.2 Rules and Connectionism

Nevertheless, it might be objected that connectionist models of language may still somehow embody linguistic rules and thus, perhaps, encompass a C/PD after all. Much has been written about the question of rules in neural networks, but for present purposes a brief overview will suffice (given the inconclusive nature of the overall discussion)[21]. Dennett (1991) argues on instrumentalist grounds that there are no rules to be found in connectionist nets, but many find that nets embody some kind of *'implicit'* rules (which is to be contrasted with the *'explicit'* rules of the classical approach). For example, Hatfield (1991) suggests that connectionist nets are rule instantiating systems (in the implicit sense of rules). Similarly, Clark (1991) notes the implicitness of rules in neural nets, stressing the need to accommodate explicit rules (perhaps by letting a net implement a virtual machine simulating a rule system, as suggested in Clark, 1989)[22]. The worry about the representation of explicit rules is echoed by Hadley (1993a) who, furthermore, strongly criticizes connectionist models for merely embodying implicit rules. In contrast, Sharkey (1992) emphasizes the flexible nature of the implicit rules in neural nets, arguing that they are well-suited to serve as soft syntactic preference rules (in contrast with their classical counterparts). Moreover, Davies (1993) has offered a distinction between two notions of implicit rules, suggesting that this distinction might help discover whether nets (and humans) embody rules.

A radically different approach has been taken by Kirsh (1990) who argues that the implicit/explicit discussion is misguided, and that we instead should take the explicitness of a rule to be dependent on whether it is accessible in constant time (for a criticism of this view and a proposal for a 'narrow' and a 'broad' sense of both implicit and explicit rules, see Hadley, 1993b). Finally, MacLennan (1991) has proposed a formalization of connectionist knowledge representation, a simulacrum (i.e., a continuous analog to discrete calculus), in which almost discrete rules can sometimes be said to emerge from network processing, whereas the latter at other times form a continuum

[21] It is worth noting that in this discussion it not always clear whether rules are implicit from *our* point of view (i.e., we cannot readily find them in the net), or from the *net's* perspective (i.e., it cannot 'see' them).

[22] In this connection, Bechtel & Abrahamsen (1991) list three possible connectionist responses to the problem of dealing with explicit rules: i) the *approximationist* approach in which the rules are considered to be mere approximations of underlying connectionist substrate; ii) the *compatibilist* view which implies that connectionist models must be able to implement the rules directly; and iii) the *externalist* position which states that networks must be able to develop the capacity to manipulate rules that are external to the net.

that is not expressible in terms of rules.

Taken together, the research on whether connectionist networks embody rules (explicit, implicit or otherwise) does not show much overall consensus. In my view, this discussion may be seen as a parallel to the above discussion of compositionality. A connectionist language model might behave *as if* it has linguistic rules, although no rules have been programmed into it. But does this warrant saying that the model has 'implicit' (or 'fuzzy') rules? Such talk about implicit rules is in danger of forcing connectionism into a symbolic mold by trying to apply a particular concept—i.e., the classical notion of a computationally efficacious rule—to connectionism. As in the case of apparent compositionality, we can only solve this problem through empirical research, developing a much better conception of what is really going on inside neural networks.

However, this still leaves the question of what kind of linguistic performance data should form the basis for connectionist models of language. For the purpose of empirical tests of candidate language mechanisms we may need to distinguish between 'real' performance data as exhibited in normal natural language behavior and examples of abnormal performance, such as, 'slips-of-the-tongue', blending errors, etc[23]. Notice that the traditional notion of 'grammaticality' does not capture the kind of data that I would want to account for. Instead, we might apply something like Reich's (1969) notion of 'acceptable' in the characterization of the data set to be modeled:

> A sentence is acceptable to me if my estimate of the probability of occurrence of a sentence of like construction in natural language is greater than zero. I exclude from natural language text sentences dreamed up by linguists, psychologists, English teachers and poets. (p. 260)

In this way, what counts as valid data is not dependent on an abstract, idealized notion of linguistic competence but on observable natural language behavior under statistically 'normal' circumstances. Consequently, we should be able operationalize Reich's notion of acceptability in order to filter out the abnormal performance data from a language corpora simply by using 'weak' statistical methods. For example, Finch & Chater (1992, 1993) applied simple distributional statistics to the analysis of a *noisy* corpus consisting of 40,000,000 English words and were able to find phrasal categories defined over similarly derived approximate syntactic categories. It seems very likely that such a method could be extended to a clausal level in order to filter out abnormal performance data.

[23]Of course, this does not mean that one should not try to model this kind of performance breakdown

Having shown the strong connection between the recursive nature of generative grammar and the competence/performance distinction, as well as providing arguments against upholding the latter distinction and in favor of a connectionist framework, which does not necessarily rely on such a distinction, I now present a number of connectionist simulations experiments in the next two chapters. In chapter 3, I present simulations demonstrating the ability of recurrent neural networks to deal with different kinds of recursion in three artificial languages involving ni-recursion, normally considered to require the power of context-free (and even context-sensitive) grammars. Chapter 4 reports the results of simulations involving a linguistic grammar of considerable complexity (incorporating sentences such as (1)–(3) and (9)–(12), that is, both i- and ni-recursive structures).

Chapter 3

Modeling Recursion in Recurrent Neural Networks

One of the Chomskyan legacies in modern linguistics is the stipulation that a finite state automaton cannot capture the linguistic structure of natural language. This position is to a large extent based on the fact that no FSA can produce strings which contain embedded substrings of arbitrary complexity and length. In this connection, the two ni-recursive linguistic phenomena—center-embedding and cross-dependency—have played an important part in the history of the theory of language. In particular, they have been used as an existence proof of the need for increasingly powerful language formalisms to describe natural language. Thus, in the second half of the fifties, Chomsky (1956, 1957) proved that regular languages cannot capture center-embedding of an arbitrary depth. Given that center-embedding does occur in natural language, he concluded that we would need at least the power of context-free (CF) languages in order to specify adequate grammars. In fact, at that time Chomsky went even further, arguing for the combination of a CF grammar and a transformational component. The latter enables transformations of CF strings into strings not derivable from the grammar itself. This was, for example, used to account for passive constructions: an active CF string, say, *'Betty loves Anita'* is transformed into the passive string *'Anita is loved by Betty'*.

About a quarter of a century later, arguments were put forward—most notably by Pullum & Gazdar (1982)—defending the sufficiency of CF grammars (without an additional transformational component). Shortly thereafter, the existence of cross-dependency in some languages was used in an attack on the adequacy of CF grammar formalisms (e.g., Shieber 1985—for a defense against that attack, see Gazdar & Pullum 1985). At the moment, this discussion is undecided, but the historic trend

within linguistics points towards increasingly powerful grammar formalisms. In contrast, connectionist and other statistically based models of natural language processing are essentially finite state machines producing 'only' regular languages. It is therefore not surprising that linguists in general do not pay much attention to these kind of approaches to natural language processing. They are dismissed *a priori* as not being powerful enough to do the job. Part of the motivation for the simulation experiments presented in this chapter (and the next) is therefore to challenge this speculative dismissal of finite state models of language.

In order to make this task more precise, consider an artificial language whose only 'words' are the letters *'a'* and *'b'*. Given this restricted basis, Chomsky (1957: p. 21) constructed the following three languages involving ni-recursion:

i. *ab, aabb, aaabbb,* ..., and in general, all sentences consisting of n occurrences of a followed by n occurrences of b and only these;

ii. *aa, bb, abba, baab, aaaa, bbbb, aabbaa, abbbba,* ..., and in general, all sentences consisting of a string X followed by the 'mirror image' of X (i.e., X in reverse), and only these;

iii. *aa, bb, abab, baba, aaaa, bbbb, aabaab, abbabb,* ..., and in general, all sentences consisting of a string X of a's and b's followed by the identical string X, and only these.

If it is required that the strings of these three languages are to be of a possibly infinite complexity, then an FSA will not suffice (even though it can produce i-recursive strings of an arbitrary length as pointed out in chapter 2). Assuming, as we must, that all parameters have finite precision, any finite neural network is also an FSA. Thus, connectionist accounts of language processing appear to inherit the shortcomings of the FSA. On the other hand, if we only need to account for a certain depth of embedding, then we can devise an FSA that can process all the strings up to that particular limit. Similarly, we might be able to train a neural network to process bounded recursion.

In chapter 2, I argued against using a distinction between an idealized competence and the actual human performance in psychological theories of natural language processing. Empirical studies (e.g., Bach, Brown & Marslen-Wilson, 1986) have shown that sentences involving three or more embeddings (such as, ii) or three or more cross-dependencies (such as, iii) are universally hard to process and understand. Nevertheless, we do need to account for a limited depth of ni-recursive embedding. The question is therefore whether this relatively simpler problem can be solved by a system adapting to mere statistical contingencies in the input data. Furthermore, is it possible for such

a system to develop an ability to generalize to new instances? In other words, can a neural network capture the ni-recursive regularities of natural language, while accepting that arbitrarily complex sentences cannot be handled? And very importantly, will the system develop behavior comparable with humans on similar tasks?

To address the above questions I have developed a number of benchmark tests based on Chomsky's three abstract languages. This chapter reports the results of these tests based on computer simulations involving two kinds of artificial neural networks as well as a program implementing n-gram statistics. The first section describes the three artificial languages that comprises the basis of the benchmark tests—both with respect to their relationship to natural language and performance expectations. Section 2 describes the neural network architectures and the n-gram statistical program. It also provides a description of the form and configuration of the input data. The next two sections report the results obtained in two experiments involving, respectively, a two word and an eight word vocabulary. The final section summarizes and discusses the results from the two benchmark experiments, comparing them with experimentally observed limitations on human processing of ni-recursive structures.

3.1 Three Bench Mark Tests Concerning Recursion

The issue of ni-recursion has been addressed before within a connectionist framework. For example, both Elman (1991a) and Servan-Schreiber, Cleeremans & McClelland (1989, 1991) have demonstrated the ability of Simple Recurrent Networks (SRN) to deal with right-branching i-recursive structures as well as limited instances of center-embedded ni-recursion (Chomsky's second abstract language). In addition, the latter form of ni-recursion has been studied further by Weckerly & Elman (1992). However, no study has directly addressed the Chomskyan challenge expressed in terms of the three abstract languages mentioned above. This chapter takes up that challenge from a connectionist perspective incorporating psycholinguistic considerations.

In order to make the tests slightly more natural language-like a constraint on Chomsky's original languages was introduced[1]. This constraint consists of an *agreement* between a 'noun' class of words and a 'verb' class of words. Each word has two forms—a lower case form and a upper case form—which can be seen as corresponding to the singular and plural forms of English. In this way, the constraint enforces an agreement

[1]I am aware that the test languages are still far from being anything like natural language. Nevertheless, they express the abstract structure of certain linguistic constraints (see below) and are therefore well-suited as benchmark tests concerning learnability and generalization. Moreover, the simulations presented in the next chapter involves more language-like data. Chapter 4 will also deals with the combination of ni- and i-recursion.

between the 'subject noun' and the verb so that lower case nouns only occur with lower case verbs and similar for upper case words (except in the case of the first language). Furthermore, each sentence has an end of sentence marker, a '.' (full stop), after the last verb[2]. The three test languages have the following structure:

i. *az., aZ., AZ., Az., aAZZ., aAazzZ., ...*, and in general, all sentences consisting of n occurrences of nouns followed by n occurrences of verbs and only these, as well as a '.'. That is, in the simple case, a combination of a's and/or A's is followed by the exact same number of z's and/or Z's plus a '.', but with *no* agreement between nouns and verbs".

ii. *az., AZ., aazz., aAZz., AAZZ., AazZ., AaazzZ., aaAZzz., ...*, and in general, all sentences consisting of a string n of nouns followed by a string n of verbs whose agreement features constitute a 'mirror image' of those in N (i.e., the order reversed with respect to N), and only these, as well as a '.'. That is, in the simple case, a combination of a's and/or A's is followed by the exact same number of z's and/or Z's plus a '.', but with the constraint that the last verb agrees with the first noun, the second but last verb with the second noun, and so forth for all words in a given sentence.

iii. *az., AZ., aazz., aAzZ., AAZZ., AaZz., AaaZzz., aaAZzz., ...*, and in general, all sentences consisting of a string n of nouns followed by a string n of verbs whose agreement features are ordered identically to those in N, and only these, as well as a '.'. That is, in the simple case, a combination of a's and/or A's is followed by the exact same number of z's and/or Z's plus a '.', but with the constraint that the first verb agrees with the first noun, the second verb with the second noun, and so forth for all words in a given sentence.

Chomsky (1957) has claimed that the first test language correspond to naturally occurring sentence constructions in English, such as, *'If* S_1*, then* S_2*'* and *'Either* S_3*, or* S_4*'* (where S_1, S_2, S_3 and S_4 are declarative sentences). These sentence structures, supposedly, can be nested arbitrarily deeply within each other as, for example, in the sentence *'If, either, if the cat is out, then let the dog in, or, the bird is out, then go for lunch'*. It should be clear from this example that sentences with only two levels of nesting become very difficult to process. Although there is no direct experimental evidence to corroborate this, Reich 1969 has suggested that these constructions only will occur naturally with a limited depth of nesting.

[2]Another motivation for the agreement constraint was to provide the systems with minimal help (e.g., cf. Cleeremans, Servan-Schreiber & McClelland, 1991).

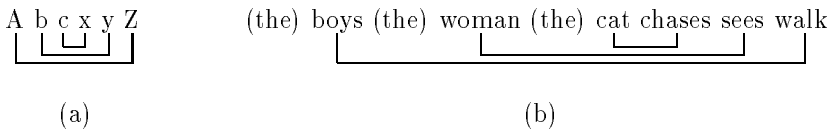
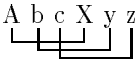


Figure 3.1. Illustration of the structural similarity between the second test language and English.

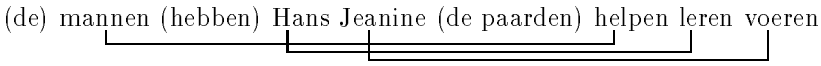
The second language corresponds structurally to *center-embedded* sentences which are found in many languages (although typically not with a high frequency). Figure 3.1(a) illustrates the subject-noun/verb dependencies in a sentence, *'AbcxyZ'*, from the second test language. These dependencies correspond structurally to the abstract structure of the subject-noun/verb agreements in (b) the English sentence, *'(the) boys (the) woman (the) cat chases sees walk'* (where lower case denotes singular and upper case plural). Again, it is clear that even sentences with two center-embeddings are difficult to process, an intuition which is supported by psycholinguistic studies (e.g., Bach, Brown, & Marslen-Wilson, 1986; Blaubergs & Braine, 1974; Larkin & Burns, 1977).

In much the same way, the structural properties of the third test language is similar to the much less common *cross-dependency* structures found in languages such as Dutch and Swiss-German[3]. Figure 3.2 shows (a) a sentence from the third test language, *'AbcXyz'*, that can be seen as structurally equivalent (in terms of subject-noun/verb dependencies) to (b) the Dutch sentence *'(de) mannen (hebben) Hans Jeanine (de paarden) helpen leren voeren'* (again with the convention that lower case denotes singular and upper case plural). The literal English translation of the latter is 'the men have Hans Jeanine the horses help teach feed' and can be glossed as 'the men helped Hans teach Jeanine to feed the horses'. As with center-embedding, there is experimental evidence suggesting a limit on the number of cross-dependencies acceptable to native speakers of Dutch (Bach, Brown, & Marslen-Wilson, 1986).

[3]The *'respectively'* constructions in English—e.g., *'Anita and Betty walked and skipped, respectively'*—are sometimes said to involve cross-dependency. However, many find that these constructions rely on semantic rather than syntactic constraints (e.g., Church 1982). Although I tend to lean towards the latter position, it will not matter for the arguments presented here if the former is adopted. In both cases there seems to be a limit on the length of acceptable 'respectively' constructions. For instance, *'Betty, the dogs, and Anita runs, walk, and skips, respectively'* is questionable at best. In unison with Church (1982), I find that three cross-dependencies in these constructions is the limit in terms of acceptability.



(a)



(b)

Figure 3.2. Illustration of the structural similarity between the third test language and Dutch.

3.1.1 Performance Expectations

When considering the three languages certain predictions can be made concerning the difficulty of processing each of them. The task in question is to predict the next word in a sentence or a sentence ending. That is, the system gets a word as input at time t and then has to predict the next word at time $t+1$. In this connection it should be noted that it is not possible to determine precisely how many nouns are going to be encountered in a sentence or what their (upper/lower case) form might be.[4] It is, however, possible to determine exactly how many verbs a sentence will have and which (upper/lower case) form each of them will have (except, of course, for the first language). Assuming that the system can learn to distinguish nouns from verbs, it should be able to make correct predictions about subsequent verbs as well as the sentence ending once it receives the first verb as input. More specifically, the number of verbs will correspond to the number of nouns, and the form of the former will agree with the form of the latter as specified by each particular language. The end of sentence marker should be predicted after the last verb.

Bearing this in mind, it appears to be the case that the first language should be the least difficult to process. Since this language does not have any agreement constraints imposed on it, a system 'merely' needs to predict the correct number of verbs in a sentence and the end of sentence marker. To perform this task the system therefore needs to count the exact number of nouns encountered. It is then able to predict

[4]Still, a learning system might develop some sensitivity to sentence length based on the length distribution of all the sentences in the input. In this way, the system would tend to predict $L/2$ nouns, where L is the average sentence length in the input (measured in words). This sensitivity will probably also cause the system to activate the end of sentence marker after it has encountered L words. For

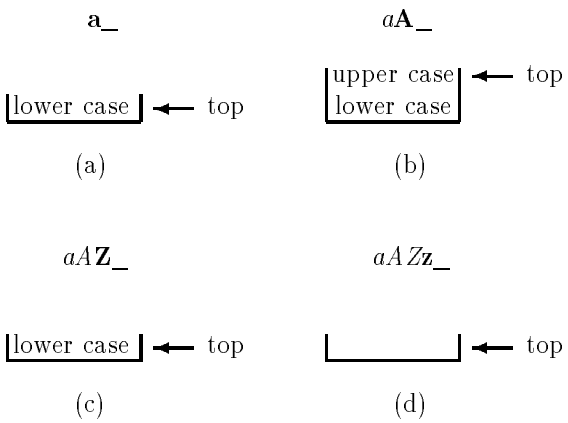


Figure 3.3. Stack contents when processing the the mirror recursive sentence *aAZz.*

the correct number of verbs as well as the end of sentence marker by decreasing its counter for each verb it predicts. For that reason I will refer to the structure of the first test language as *counting recursion.* As an example, imagine that the system previously has received *aA***Z**_[5] as its input. If the system has counted the number of nouns encountered it should be able to predict the next word as being a verb (of either form). Next, it should predict the end of sentence marker as the last 'word'.

The most efficient way to process the second language is to develop a stack-like memory structure (as we saw in chapter 2); that is, to 'implement' a last-in-first-out memory storing the agreement forms of the nouns in the exact order they were encountered. Once the system receives the first verb as input, its form should agree with the noun form on the top of the stack. This form information is removed from the stack. The next verb form can then be predicted as corresponding to the noun form which constitutes the new stack top. The same procedure is followed for subsequent verb form predictions until the stack is empty and an end of sentence marker can be predicted. Using this memory structure a system will be able to predict the agreement form of all verbs (save the first) as an exact mirror image of the noun forms. I will therefore refer to the structure of the second test language as *mirror* recursion.

Consider as an example the mirror recursive sentence *'aAZz.'*. After having received the first noun, *'a'*, as input, the stack would look something like Figure 3.3(a);

sentences longer than this average this might lead to increasingly inaccurate predictions.

[5]I will adopt the following typographical convention regarding prediction context: All the previous input words will be listed up to and including the current input word. The latter will be emphasized in bold. Underscore signifies the item to be predicted.

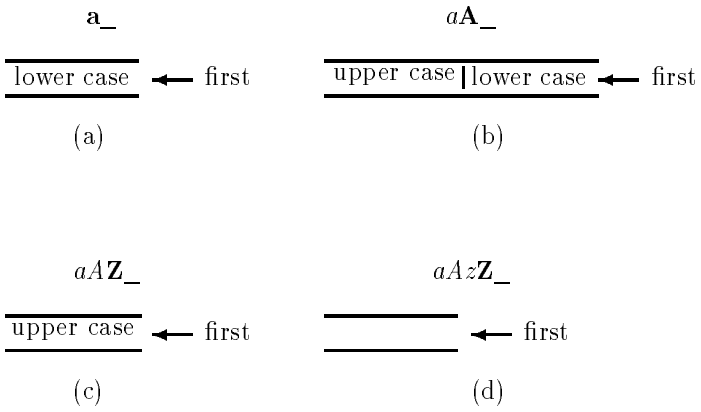


Figure 3.4. Queue contents when processing the identity recursive sentence *aAzZ.*

indicating that if the next word is a verb it must be lower case. Since the next word is another 'noun', its dependency constraint is put *on top* of the stack as illustrated in (b), predicting that if a verb is next it should be upper case. When the system gets the first verb, as in (c), the agreement form on the top of the stack is removed. The new top (lower case) is used to predict the form of the next word as a lower case verb. After receiving the last verb the stack becomes empty (d), and an end of sentence marker can be predicted.

To process the third language most efficiently a system needs to develop a queue-like memory structure. This is a first-in-first-out memory where the agreement forms of the nouns are stored in the exact order they are received as input (while keeping track of both first and last word form). Once the first verb is encountered its form will agree with the first noun form in the queue. This word form is removed from the queue and the next verb form can then be predicted as having the form of the new head of the queue. Subsequent verb forms can be predicted using the same procedure. When the queue becomes empty the end of sentence marker should be predicted. By using a queue memory structure a system is able to predict the agreement forms of all the verbs (save the first) as being identical to those of the encountered nouns. Consequently, I will refer to the structure of the third test language as *identity* recursion.

As an example consider the identity recursive sentence *'aAzZ.'*. Figure 3.4 illustrates the queue states while processing this sentence. Having received a noun, *'a'*, as input, its dependency constraint is stored in the queue (a). This allows the system to predict that if the next word is a verb it should have a lower case form. In (b), another noun, *'A'*, has been given as input, and its agreement form stored *behind* the previous

form in the queue. Thus, a lower case form is still predicted as the first occurring verb. When the system gets the first verb, *'z'*, the first element of the queue is removed. This leaves the upper case form in the queue (c), indicating that the next verb must have this agreement form. Finally, the queue becomes empty after the system receives the second verb, *'Z'*, and an end of sentence marker can be predicted (d).

At first glance the two last test languages appear to be equally hard in terms of making predictions about the form of the verbs. Nonetheless, I suggest that mirror recursion should be easier to process than identity recursion. This is because the former appears to be less demanding in relation to both memory load and learning. First, in traditional (symbolic) implementations of stacks only one pointer is needed to keep track of the stack top if it is assumed that the individual elements are linked together by pointers. In contrast, given the same assumption two pointers are necessary to build and maintain a queue structure (pointing to the first and last element, respectively). This creates a higher load on a system's memory, which, in turn, might impair performance. Secondly, there is a fundamental difference between the two languages with respect to learning. In the case of mirror recursion, strings with an embedding depth, D, generalizes directly to the next depth of embedding, $D + 1$. For example, the simple string *az* generalizes to the more complex strings *aazz* and *AazZ*. These, in turn, generalize to strings at the next embedding depth: *AAazZZ, aAazZz, AaazzZ,* and *aaazzz*, and so on *ad infinitum*. This is, however, not the case for identity recursion. In this language, the most simple strings, e.g., *az*, *do not* generalize directly to the more complex strings such as *AaZz* or *AAaZZz*. That is, the system cannot use generalizations from simple strings to facilitate the learning of more complex strings. This is very likely to make learning of identity recursive structures more difficult.

In addition, I expect there to be differences between the second and third test languages in the way the length of a sentence will influence prediction accuracy[6]. It is reasonable to assume that a system only has a limited amount of memory available for the implementation of a stack or a queue structure. This, in turn, limits the size, S, of the stack or queue (measured in terms of the number of elements they can maximally contain). Given S and the length, L, of a sentence we can envisage prediction problems when $L/2$ becomes greater than S. Regarding mirror recursion, the system should

[6]Note that if a system implements the most efficient way of processing counting recursion—that is, develop a counter—then the length of a sentence should not matter. However, it is likely that neural networks are not able to adopt this strategy. For example, Servan-Schreiber, Cleeremans & McClelland (1991) report simulations which suggest that SRNs tend to develop stack-like memory structures. In relation to counting recursion this implies that the system would store redundant information about previous input. If this is the case, then the system should exhibit the same behavior as systems dealing with mirror recursion—but perhaps with slightly better performance.

be able to make fairly accurate predictions with respect to the S innermost symbols independent of L.[7] This is because these symbols will always be on the top part of the stack. For example, consider the sentence $aAAaAZzZZz$ for S being 3. Although, the system might lose agreement information about the two first nouns, a and A, it should still be able to make fairly accurate predictions about the second and the third verbs. In effect, the system is likely to 'see' the sentence as $AaazzZ$ and predict an end of sentence marker after the third verb. On the other hand, prediction performance on identity recursion is likely to break down altogether when $L/2$ exceeds S because the front of the queue will be lost. Consider the sentence $aAAaAzZZzZ$ as an example for S being 3. Since it has lost the two first noun forms, a and A, it will erroneously expect the first verb to be a Z because its queue will only contain agreement information about AaA. The prediction for the next verb will also be wrong (z instead of Z) whereas the prediction for the third verb form will be correct by chance (Z). Overall this should lead to better performance on the second test language than on the third.

3.2 Network Architectures and n-gram Stats

One way of approaching the problem of dealing with recursion in connectionist models is to "hardwire" symbolic structures directly into the architecture of the network. Much early work in connectionist natural language processing (e.g., McClelland & Kawamoto, 1986) adopted this implementational approach. Such connectionist re-implementations of symbolic systems might have interesting computational properties and even be illuminating regarding the appropriateness of a particular style of symbolic model for distributed computation (Chater & Oaksford, 1990a). On the other hand, there is the promise that connectionism may be able to do more than simply implement symbolic representations and processes; in particular, that networks may be able to *learn* to form and use structured representations. The most interesting models of this sort typically focus on learning quite constrained aspects of natural language syntax. These models can be divided into two classes, depending on whether preprocessed sentence structures or simply bare sentences are presented.

The less radical class (e.g., Hanson & Kegl, 1987; Pollack, 1988, 1990; Sopena, 1991; Stolke, 1991) presupposes that the syntactic structure of each sentence to be learned is

[7] In fact, since neural networks only approximate a traditional stack structure (cf. Servan-Schreiber, Cleeremans & McClelland, 1991; Weckerly & Elman, 1992) it is to be expected that prediction accuracy will deteriorate gradually as we move down the stack (even within S). Performance on identity sentences where S is greater than $L/2$ should likewise exhibit the same graded behavior.

given. The task of the network is to find the grammar which fits these example structures. This means that the structural aspects of language are not themselves learned by observation, but are built in. These models are related to statistical approaches to language learning such as stochastic context free grammars (Brill *et al.*, 1990; Jelinek, Lafferty, & Mercer, 1990) in which learning sets the probabilities of each grammar rule in a prespecified context-free grammar, from a corpus of parsed sentences.

The more radical models have taken on a much harder task, that of learning syntactic structure from strings of words, with no prior assumptions about the particular structure of the grammar. The most influential approach is to train SRNs developed by Jordan (1986) and Elman (1988). These networks provide a powerful tool with which to model the learning of many aspects of linguistic structure (for example, Cottrell & Plunkett, 1991; Elman, 1990, 1991a; Norris, 1990; Shillcock, Levy & Chater, 1991); there has also been some exploration of their computational properties (Chater, 1989; Chater & Conkey, 1992; Cleeremans, Servan-Schreiber & McClelland, 1989; Maskara & Noetzel, 1992, 1993; Servan-Schreiber, Cleeremans & McClelland, 1989, 1991). It is fair to say that these radical models have so far reached only a modest level of performance. In general, it seems to be possible to learn simple finite state grammars involving left and right recursion. Still, only little headway has been made towards more complex grammars involving center-embedded recursion (most noticeable by Elman, 1991 and Weckerly & Elman, 1992—furthermore, see next chapter for a non-trivial extension of these results), but not towards cross-dependency recursion. The simulations reported in this chapter build on and extend this work.

The SRN is a limited version of a more general neural network architecture: *viz.,* a fully recurrent network. Figure 3.5(a) shows an example of such a network. These networks are, however, difficult to train because they involve lateral connections between the units in the hidden layer, thus preventing the application of standard back-propagation learning. Recurrent networks are therefore usually trained by 'unfolding" them into feedforward networks with the same behavior. The hidden units from the previous time-step are then treated as an additional set of inputs, allowing the resulting feedforward network to be trained using conventional back-propagation.

There are various ways in which this unfolding can be achieved (see Chater & Conkey, 1992). One approach is to unfold the network through several time steps (Rumelhart, Hinton & Williams, 1986) so that each weight has several 'virtual incarnations' and then back-propagate error through the resulting network. The overall weight change is simply the sum of the changes recommended for each incarnation. As illustrated in Figure 3.5(b), this 'back-propagation through time'—or, Recurrent Back-Propagation (RBP)—is typically implemented by unfolding through a small number of

time steps (7 for the current simulations).

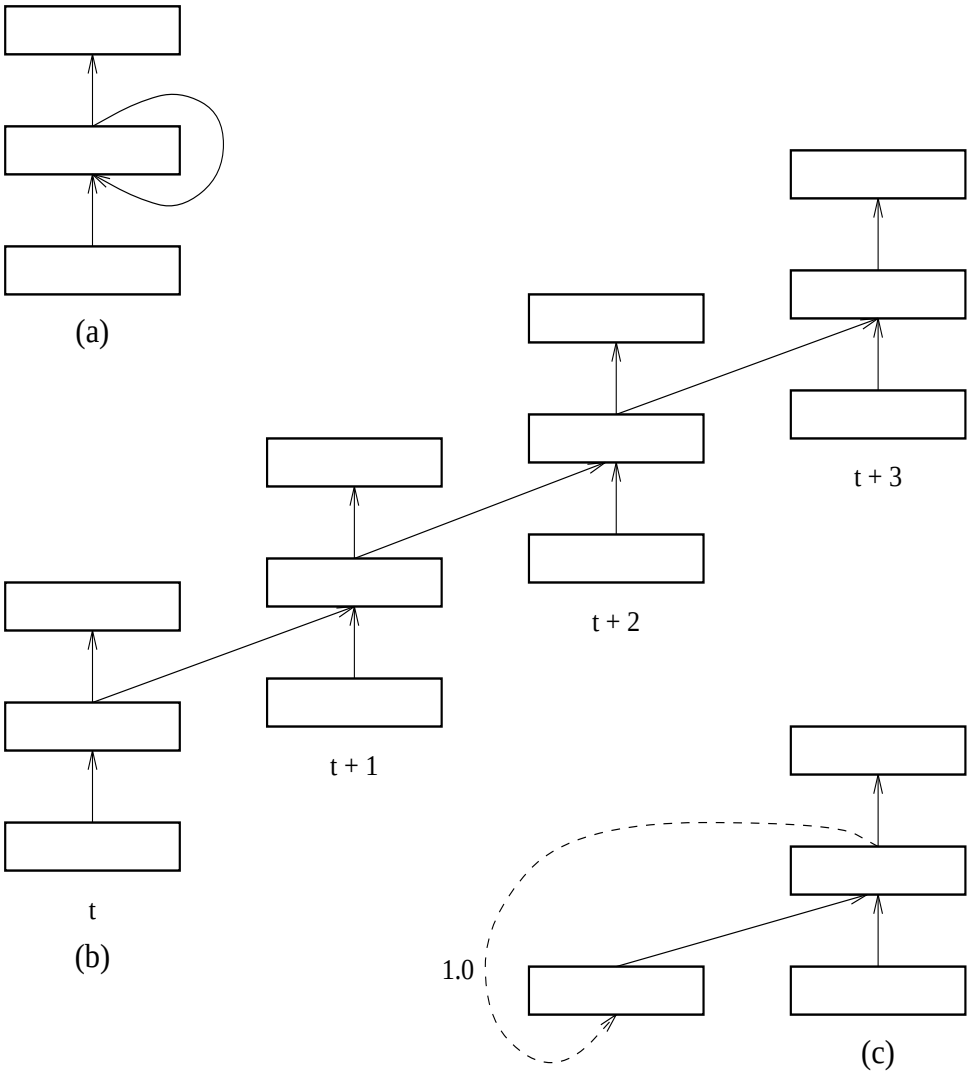


Figure 3.5. Unfolding a recurrent neural network (a) into a back-propagation through time network (b) and into a simple recurrent network employing copy-back training (c).

The copy-back scheme employed in SRNs can be viewed as a special case of RBP, in which the back-propagation of error stops at the first copy of the hidden units—the context units—as indicated Figure 3.5(c). Note that the one-to-one copy-back links going from the hidden layer to the context layer always have a value of 1.0, copying the activation of the hidden units at time t to the context units so they can be used as input at $t+1$. Simulations by Chater & Conkey (1992) have shown that RBP performs better than SRNs on a number of tasks (such as, learning to be a delay line and performing discrete XOR), although the former is considerably more computationally expensive.

A secondary motivation for the present simulations is therefore to compare the two training regimes on tasks that more closely resemble language[8].

In order to provide an independent basis for assessing the performance of the two kinds of networks, I developed a simple statistical prediction method, based on n-grams, strings of n consecutive words. The program is "trained" on the same stimuli used by the networks, and simply records the frequency of each n-gram in a look-up table. It makes predictions for new material by considering the relative frequencies of the n-grams which are consistent with the previous $n-1$ words. The prediction is a vector of relative frequencies for each possible successor item, scaled to sum to 1, so that they can be interpreted as probabilities, and are therefore directly comparable with the output vectors produced by the networks. Below, I report the predictions of bigram, quadrogram, hexagram and octogram models and compare them with network performance.

3.3 Experiment 1: Two Word Vocabulary

The first experiment involves the vocabulary found in Chomsky's description of his three abstract languages (with the additional singular/plural agreement constraint as described in section 3.1). Thus, we have a vocabulary consisting of a noun in a singular and plural form, *'a'* and *'A'* respectively, and a verb likewise in a singular and plural form, *'z'* and 'Z' respectively. Each of the networks in experiment 1 was trained with 5, 10 and 25 hidden units on a data set consisting of 2000 sentences of variable length. For each sentence the depth of nesting was computed by iterating the following: if $r < p^n(1-p)$ then an extra level of nesting would be added to the sentence, where r is a random number between 0 and 1; p the probability of adding a level of nesting (0.3 in the simulations reported here); and n the number of nestings that the sentence already has. Then all the nets in the present experiment were tested on a data set consisting of 1000 sentences, generated in the same way as the training set. The inputs and output were represented as binary localist vectors with one bit for each word form and one for the end of sentence marker (totaling 5 inputs/outputs).

Initial explorations indicated that the best performance for the SRNs was to be obtained with a learning rate of 0.5, a momentum of 0.25 and an initial randomization

[8]In any interesting language-like task, the next item will not be deterministically specified by the previous items, and hence it is appropriate for the prediction to take the form of a probability distribution of possible next items. Consequently, network performance in the simulations reported below was measured against this probability distribution directly, rather than against predictions of the specific next item in the sequence. Following Elman (1991) the mean cosine between network output vectors and probability vectors given previous context is used as the main quantitative measure of performance.

of the weights between ± 0.5. In the case of RBP, no momentum was used, the learning rate was set to 0.5 and the weights initialized randomly between ± 1.0. Through cross-validation it was found that the number of epochs necessary to reach peak performance in both cases varied with the size of the hidden unit layer. Increasing the hidden unit layer resulted in faster training (although the RBP net exhibited much faster training across the board)[9]. Subsequently, the SRNs with 5, 10 and 25 hidden units were trained for 500, 450 and 350 epochs, respectively. The RBP network with 5, 10 and 25 hidden units were trained for 275, 250 and 200 epochs, respectively.

3.3.1 Counting Recursion

The networks were trained on a training set consisting entirely of sentences with counting recursive structures of variable length (mean: 4.69; sd: ± 1.37) and then tested on a test set (mean: 4.76; sd: ± 1.36). Table 3.1 shows the embedding distribution in both data sets.

Embedding	Depth 0	Depth 1	Depth 2	Depth 3
Training set	31.15%	54.10%	13.65%	1.10%
Test set	29.20%	54.40%	15.70%	0.70%

Table 3.1. The distribution of different depths of embedding in the training and test sets involving counting recursion.

General Performance

Both nets generally performed well on the counting recursion task[10]. The simulation results are reported in table 3.2. It is clear that the nets have picked up the sequential structure of the data, otherwise their performance would not have surpassed the level of performance obtained by merely making predictions according to the simple relative frequency of the words in the data set (1gram)[11]. In fact, the performance of both nets is at the same level as the bigram based performance. This could suggest that net processing is sensitive to the bigram statistics found in the input data. However, the nets are not able to perform as well as quadrogram, hexagram and octogram based

[9] However, even though the SRNs require more epochs to learn a task, they are faster in overall computing time, because the RBP nets are very expensive in computational terms and take more CPU time per epoch.

[10] This is in comparison with Elman's (1991) results. He reported a mean squared error of 0.177 and a mean cosine of 0.852. Perfect performance would have resulted in 0.0 and 1.0, respectively.

[11] This level of performance (1gram) is what, at best, could be expected from a standard feedforward back-propagation network without *any* recurrent connections.

Network/ n-gram		Mean squared error	Mean Cosine
srn	5hu	0.2060 ± 0.1841	0.8143 ± 0.1426
	10hu	0.1817 ± 0.1938	0.8450 ± 0.1258
	25hu	0.2315 ± 0.2214	0.8095 ± 0.1573
rbp	5hu	0.2145 ± 0.2285	0.8469 ± 0.1548
	10hu	0.2862 ± 0.1579	0.7746 ± 0.1279
	25hu	0.1758 ± 0.2519	0.8636 ± 0.1618
1gram		0.3464 ± 0.2468	0.6481 ± 0.1226
2gram		0.1876 ± 0.1904	0.8301 ± 0.1705
4gram		0.0570 ± 0.2120	0.9589 ± 0.1558
6gram		0.0301 ± 0.1163	0.9724 ± 0.0802
8gram		0.1026 ± 0.2081	0.9042 ± 0.1905

Table 3.2. General performance on counting recursive structures.

performance. Also notice the decrease in performance for the octogram based predictions. This is presumably caused by the limited size of the training set which leads to too many single occurrences of unique octograms.

Network/		Mean squared error	Mean Cosine
srn	5hu	0.9506 ± 0.4326	0.6515 ± 0.2132
	10hu	1.0844 ± 0.1402	0.6214 ± 0.0934
	25hu	0.4008 ± 0.2347	0.6441 ± 0.1147
rbp	5hu	1.4760 ± 0.2458	0.5956 ± 0.1283
	10hu	1.4055 ± 0.1956	0.6213 ± 0.0984
	25hu	1.5039 ± 0.5310	0.6235 ± 0.2188

Table 3.3. Baselines for the general performance on counting recursive structures.

Table 3.3 shows net performance on the test set measured before any learning had taken place. Mean squared error is high for all net configurations (except for the SRN with 25 hidden units[12]). The mean cosines are all near the performance found by predicting according to the simple relative frequency of the words in the data set (1gram). This suggests that the net configurations are well-suited for the task at hand. Still, learning improves performance across the board by approximately 30%.

[12]The low mean squared error for the SRN with 25 hidden units was also found for the same net in the mirror and identity recursion task because all nets of the same size started out with the same set of initial weights for the sake of cross-task comparisons. However, the mean cosines for these SRN configurations are of the approximately same size as the ones for the other two configurations. Since mean cosines are going to be the quantitative focus of the performance assessments, the present mean squared error anomaly should be of no importance.

Network/ n-gram		Mean Luce ratio difference	Mean Luce ratio hits	Mean Luce ratio misses
srn	5hu	0.2967 ± 0.1904	0.3780 ± 0.0689	0.3769 ± 0.0645
	10hu	0.2635 ± 0.2034	0.3601 ± 0.0639	0.3578 ± 0.0646
	25hu	0.2825 ± 0.2103	0.4361 ± 0.0561	0.4385 ± 0.0542
rbp	5hu	0.2626 ± 0.2225	0.4882 ± 0.0785	0.5163 ± 0.0933
	10hu	0.3519 ± 0.1474	0.6137 ± 0.1400	0.6427 ± 0.1660
	25hu	0.1699 ± 0.1716	0.5508 ± 0.1427	0.4769 ± 0.1073
2gram		0.2309 ± 0.1760	0.4460 ± 0.1234	0.4307 ± 0.1158
4gram		0.0832 ± 0.1359	0.5657 ± 0.2100	0.5559 ± 0.1115
6gram		0.0671 ± 0.1018	0.6308 ± 0.2499	0.5304 ± 0.1069
8gram		0.1377 ± 0.1831	0.6956 ± 0.2475	0.5734 ± 0.2435

Table 3.4. Confidence in predictions concerning counting recursive structures.

The amount of confidence that the nets had in their predictions is shown in table 3.4, measured as the Luce ratio (that is, the activation of the most active unit in proportion to the sum of the activation of the remaining units). The mean Luce ratio difference provides a measure of the absolute difference in confidence between a set of predictions and the full conditional probabilities. The RBP network generally seems to be slightly closer than the SRN to the full conditional probabilities in terms of confidence. This is also mirrored in the higher Luce ratio for both hits and misses[13] obtained by the former. Notice, however, that whereas the SRNs have the same confidence regarding both hits and misses, the RBP nets have a higher confidence in their misses (except the net with 25 hidden units which is the net that did best on all measures).

Comparisons with the n-gram predictions again indicate that net performances is on the level of bigram predictions (though the RBP net does somewhat better). Mean Luce ratio differences are exceptionally low for quadrogram and hexagram predictions, and only slightly higher for octogram predictions. Predictions based on these three n-grams also exhibit a higher confidence in their hits, compared with their misses.

Embedding Depth Performance

Since we are dealing with recursive structures of variable length it is worthwhile looking at performance across recursive embedding depth. Figure 3.6 illustrates that network performance generally decreases across embedding depth (except for the RBP network with 5 hidden units—its behavior might be a product of slight overtraining). This

[13] A hit is recorded when the highest activated unit in a prediction vector is also the highest activated unit in the full conditional probability vector. When this is not the case a miss is recorded.

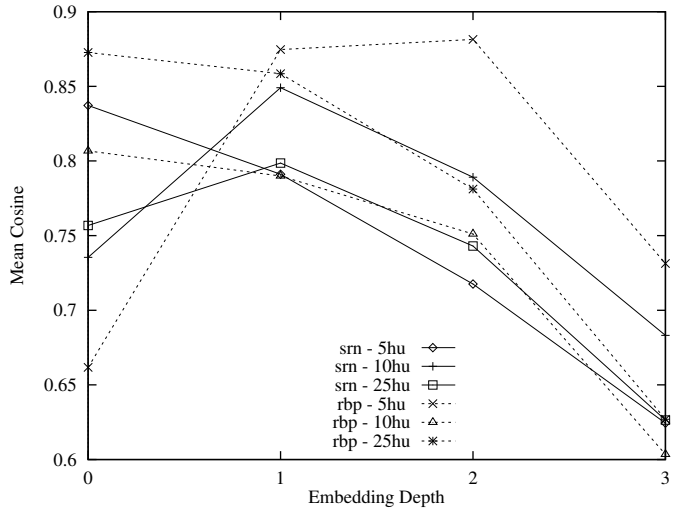


Figure 3.6. Network performances on counting recursion plotted against embedding depth.

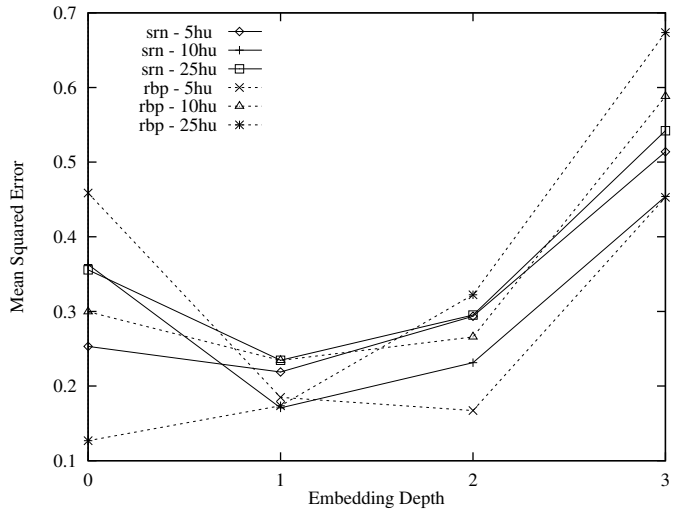


Figure 3.7. Network error regarding counting recursive structures plotted as a function of embedding depth.

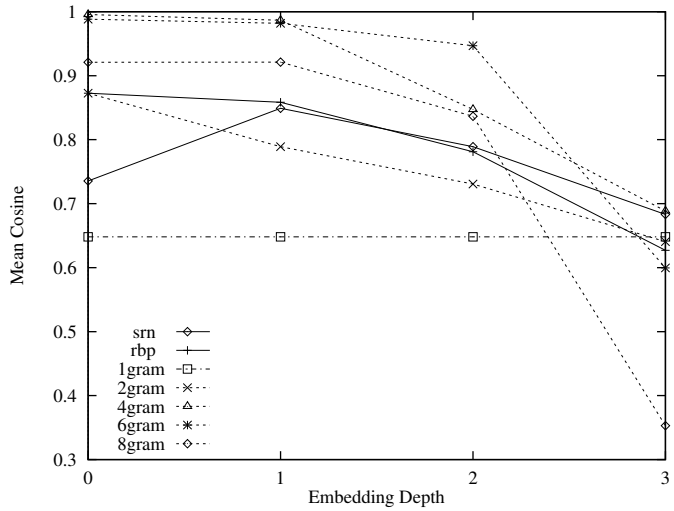


Figure 3.8. Network (SRN: 10hu; RBP: 25hu) and n-gram performance regarding counting recursion plotted against embedding depth.

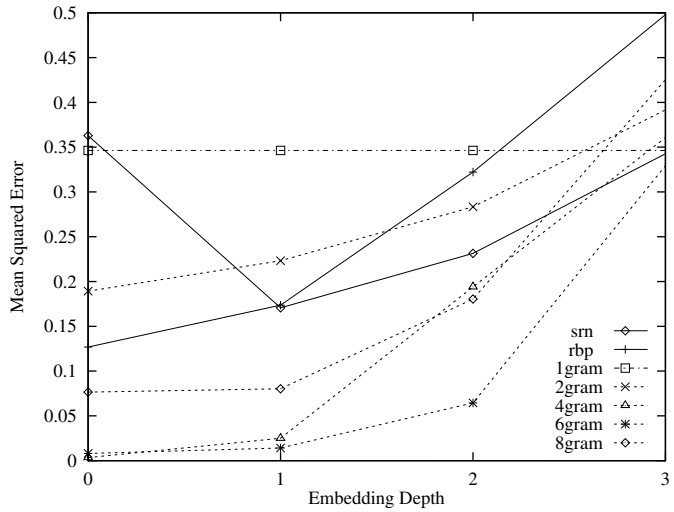


Figure 3.9. Network (SRN: 10hu; RBP: 25hu) and n-gram error regarding counting recursive structures plotted as a function of embedding depth.

is exactly what we would expect according to the performance predictions made in Section 3.1.1. Figure 3.7 shows the same network behavior, this time measured in terms of mean squared error.

From Figure 3.8 we can, once again, see that the performances of the best net of each kind (the SRN with 10 hidden units and RBP network with 25 hidden units) has a performance across recursion depth that is comparable with that of the bigram predictions (perhaps even slightly better). The higher n-gram predictions are once more superior than the net predictions. It is worth noticing that even though the accuracy of net and n-gram predictions degrades over time, all predictions seem to meet at depth three at the level of predictions discernible using only the simple relative frequence of the words in the data set (except the octogram predictions which for the reason mentioned above deteriorates rather than degrade). Figure 3.9 illustrates the same net and n-gram prediction behavior, but measured in terms of mean squared error (since the mean squared error graphs essentially are mere mirror images of the mean cosine graphs, only the latter will be presented hereafter).

In summary, the counting recursion simulations came out very much as expected with performance gracefully degrading over embedding depth. However, the difference in the size of the hidden unit layer did not seem to matter much given the task at hand. In addition, it seems to be the case that both kinds of nets can only learn to be sensitive to bigram stats.

3.3.2 Mirror Recursion

In this task the nets were trained on a training set containing only sentence structures of variable length from the mirror recursive test language (mean: 4.73; sd: $\pm$1.36). After training the nets were tested on a different mirror recursion data set (mean: 4.79; sd: $\pm$1.33). The distribution of the sentence embeddings is shown in Table 3.5.

Embedding	Depth 0	Depth 1	Depth 2	Depth 3
Training set	29.55%	55.50%	13.80%	1.15%
Test set	26.90%	57.50%	14.70%	0.90%

Table 3.5. The distribution of different depths of embedding in the training and test sets involving mirror recursion.

General Performance

The performance of the two nets on this task is generally poorer than on the counting recursion task, but still at an acceptably high level.[14] Moreover, this difference in

Network/ n-gram		Mean squared error	Mean Cosine
srn	5hu	0.3417 ± 0.3062	0.7508 ± 0.1874
	10hu	0.3122 ± 0.2886	0.7953 ± 0.1783
	25hu	0.3550 ± 0.4610	0.7553 ± 0.2567
rbp	5hu	0.3062 ± 0.4822	0.7914 ± 0.2674
	10hu	0.2991 ± 0.3032	0.8069 ± 0.2042
	25hu	0.3689 ± 0.4070	0.7692 ± 0.2353
1gram		0.5003 ± 0.2652	0.5688 ± 0.1161
2gram		0.2877 ± 0.3281	0.8071 ± 0.2176
4gram		0.0605 ± 0.2642	0.9660 ± 0.1585
6gram		0.0121 ± 0.0951	0.9901 ± 0.0678
8gram		0.0221 ± 0.1051	0.9784 ± 0.1092

Table 3.6. General performance on mirror recursive structures.

performance was to be expected cf. Section 3.1.1. The results—shown in Table 3.6—once again indicate that net performances are comparable with bigram predictions and that predictions made using larger n-grams are superior to net predictions. Notice that performance based on the higher n-grams are slightly better than their respective performance on the counting recursion task. In addition there is an improvement of the octogram based performance (i.e., octograms are doing better than quadrograms, whereas the opposite was the case in the previous task). The overall performance improvement obtained by predictions based on higher n-gram statistics can plausibly be ascribed to the existence of more deterministic structure in the mirror recursive task.

From Table 3.7 it can be seen that even though learning increased performance approximately 20 percentage points on both counting and mirror recursion, the relative increase in performance on mirror recursion was considerably bigger (about 45% versus 30% for counting recursion). Learning also provided a bigger increase in performance on the present task compared with the previous one, when the difference between the obtained net performance and the performance based on the simple relative frequencies of the word in the data set is considered (1gram). In addition, it should be noted that

[14] Both nets displayed a performance below what Elman (1991) has reported, but well above the performance obtained by Weckerly & Elman (1992) on mirror recursive structures.

the performance of the untrained nets is a great deal worse (about 30%) on the mirror recursion task compared with the previous task. This suggests that the present task is more difficult than counting recursion (which was also predicted in Section 3.1.1).

Network		Mean squared error	Mean Cosine
srn	5hu	1.1040 ± 0.4187	0.5675 ± 0.1934
	10hu	1.2309 ± 0.2962	0.5568 ± 0.1278
	25hu	0.5579 ± 0.2517	0.5602 ± 0.1033
rbp	5hu	1.6253 ± 0.4779	0.5381 ± 0.1923
	10hu	1.5590 ± 0.2442	0.5425 ± 0.0946
	25hu	1.6616 ± 0.6922	0.5621 ± 0.2566

Table 3.7. Baselines for the general performance on mirror recursive structures.

On the mirror recursion task, the nets were generally less confident about their predictions than on the counting recursion task. This is expressed in the higher mean Luce ratio differences found in Table 3.8. The same is the case for the bigram based predictions. On the other hand, the higher n-gram predictions are actually doing slightly better on this measure (especially octogram based predictions). This is also

Network/ n-gram		Mean Luce ratio difference	Mean Luce ratio hits	Mean Luce ratio misses
srn	5hu	0.4185 ± 0.2495	0.3666 ± 0.0810	0.3519 ± 0.0869
	10hu	0.3782 ± 0.2984	0.3998 ± 0.0615	0.3382 ± 0.0477
	25hu	0.3767 ± 0.2448	0.5343 ± 0.0928	0.4545 ± 0.1079
rbp	5hu	0.2810 ± 0.2757	0.5769 ± 0.1784	0.5747 ± 0.1020
	10hu	0.3486 ± 0.2399	0.6290 ± 0.1454	0.5255 ± 0.1158
	25hu	0.3671 ± 0.2238	0.6939 ± 0.1333	0.6364 ± 0.1113
2gram		0.3229 ± 0.2659	0.5266 ± 0.0130	0.5360 ± 0.0003
4gram		0.0744 ± 0.1697	0.7445 ± 0.2190	0.4725 ± 0.1712
6gram		0.0329 ± 0.0757	0.7980 ± 0.2309	0.4387 ± 0.1095
8gram		0.0484 ± 0.1100	0.8202 ± 0.2251	0.4513 ± 0.1403

Table 3.8. Confidence in predictions concerning mirror recursive structures.

mirrored in the distribution of mean Luce ratios on hits and misses for the higher n-grams. There is a bigger gap between the confidence exhibited on correct predictions and on incorrect ones—resulting from an increase of the mean Luce ratios for the former and a decrease with respect to the latter. Net confidence on hits and misses replicate this pattern (although not with the same significant gap between their confidence in correct and incorrect predictions). This differs from the pattern of net mean Luce

ratios on hits and misses found on the counting recursion task, suggesting that the more deterministic structure in the mirror recursion data sets allows the nets (and n-gram program) to be more confident about their predictions.

Embedding Depth Performance

Turning to performance in terms of embedding depth, Figure 3.10 illustrates that performance generally degraded over embedding depth. This is of the same pattern as found in the counting recursion task and in unison with the performance anticipations expressed in Section 3.1.1. Figure 3.11 shows the comparison between the best net performance for each net and the performance based on n-grams. Again, the performance of the two nets are comparable with the performance of bigram predictions. In comparison, the higher n-gram performances follows a different trend with a high performance up till an embedding depth of two, followed by a quite dramatical drop in performance on depth three (presumably caused by the very few occurrences (1.10%) of depth three recursive structures found in the training set). Notice that the higher n-gram performance across embedding depth is better on the present task than on the previous task (compare with Table 3.7). Moreover, the drop in octogram based performance does not force it below the performance based on the simple relative frequency of the words in the data set. In fact, neither net performance nor performance based on n-gram predictions falls below this level of performance.

To sum up, the nets generally did less well on the mirror recursion task compared with the counting recursion task. However, this was to be expected from the discussion of performance anticipations in section 3.1.1. As on the previous task, performance degrades over embedding depth which is at par with human performance on similar (center-embedded) structures. Furthermore, the nets were not able to do better than bigram based predictions on this task either.

3.3.3 Identity Recursion

The networks were trained on a data set consisting exclusively of identity recursive sentences of variable length (mean: 4.73; sd: $\pm$1.39). Then nets were subsequently tested on a separate data set (mean: 4.70; sd: $\pm$1.35) derived in the same manner as the training set. Table 3.9 shows the embedding distribution in the two data sets.

General Performance

The performance of both nets on this task were poorer than on the previous two tasks—confirming part of the analysis of the processings difficulties in Section 3.1.1.

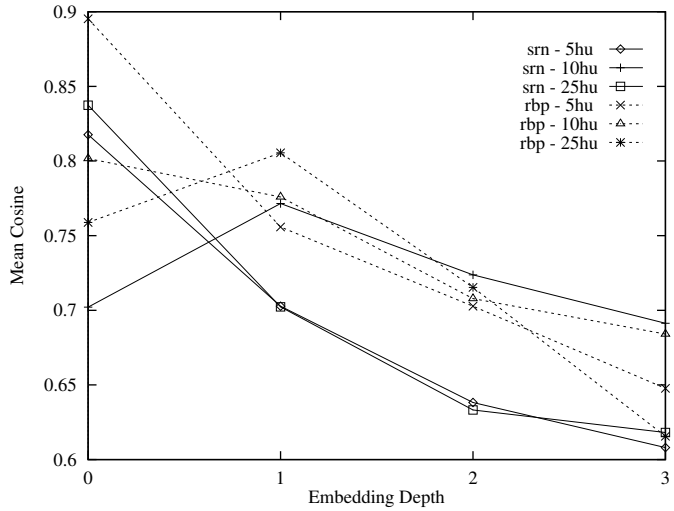


Figure 3.10. Network performance regarding mirror recursion plotted against embedding depth.

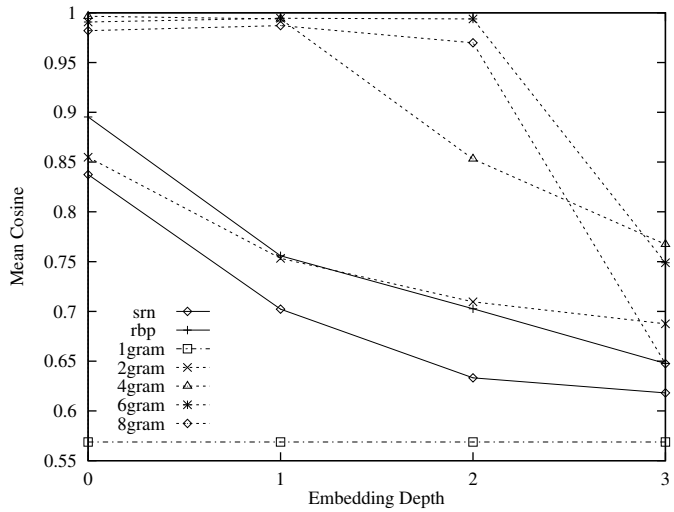


Figure 3.11. Network (SRN: 25hu; RBP: 5hu) and n-gram performance regarding mirror recursion plotted against embedding depth

Embedding	Depth 0	Depth 1	Depth 2	Depth 3
Training set	30.70%	53.10%	15.10%	1.10%
Test set	30.40%	54.80%	13.90%	0.90%

Table 3.9. The distribution of different depths of embedding in the training and test sets involving identity recursion.

However, the performance was only slightly worse on the present task than on the mirror recursion task. The nets, then, did better on the identity recursion task than was expected given the structural complexity of the learning task. In comparison

Network/ n-gram		Mean squared error	Mean Cosine
srn	5hu	0.3521 ± 0.3208	0.7367 ± 0.2150
	10hu	0.3815 ± 0.2817	0.7185 ± 0.1972
	25hu	0.3722 ± 0.3067	0.7306 ± 0.2067
rbp	5hu	0.4893 ± 0.3358	0.6696 ± 0.2245
	10hu	0.4857 ± 0.3007	0.6904 ± 0.2200
	25hu	0.3442 ± 0.3915	0.7409 ± 0.3114
1gram		0.4960 ± 0.2657	0.5714 ± 0.1178
2gram		0.3308 ± 0.3352	0.7621 ± 0.2371
4gram		0.0514 ± 0.2565	0.9717 ± 0.1427
6gram		0.0085 ± 0.0593	0.9914 ± 0.0506
8gram		0.0188 ± 0.0813	0.9811 ± 0.0853

Table 3.10. General performance on identity recursive structures.

with the predictions based on n-grams, the nets, once again, display the same level of performance as bigram predictions. This is furthermore reflected in the decrease of the bigram based performance on the identity recursion task compared with the previous task. The performance obtained by higher n-gram predictions, on the other hand, is at the same level as on the previous task (and subsequently slightly higher than on the counting recursion task).

Since the baselines as reported in Table 3.11 are similar to those found on the mirror recursion task (Table 3.7), the relative increase in performance obtained through learning is smaller on the present task than on the previous task (approximately 35% versus 45%). This indicates that identity recursion is relatively harder to learn than mirror recursion (although perhaps not as difficult as anticipated). Still, the nets were able to pick up a considerable part of the sequential structure in the training set as evinced by the gap between the performance of the two nets and that based on the simple relative word frequency.

Network		Mean squared error	Mean Cosine
srn	5hu	1.0980 ± 0.4227	0.5712 ± 0.1971
	10hu	1.2290 ± 0.2897	0.5571 ± 0.1271
	25hu	0.5476 ± 0.2515	0.5680 ± 0.1048
rbp	5hu	1.6166 ± 0.4656	0.5422 ± 0.1880
	10hu	1.5448 ± 0.2349	0.5485 ± 0.0969
	25hu	1.6383 ± 0.6744	0.5704 ± 0.2507

Table 3.11. Baselines for the general performance on identity recursive structures.

Prediction confidence as expressed in terms of the mean Luce ratio differences presented in Table 3.12 was smaller than on the mirror recursion task. This trend is also reflected in the bigram based prediction confidence, but not by the higher n-gram predictions. The latter show the same confidence as on the previous task—with a small increase in confidence for both correct and incorrect predictions. The higher confidence exhibited by the higher n-grams on correct predictions compared with incorrect predictions noted with respect to the mirror recursion task can also be found on the present task. The network confidence pattern on hits and misses are less clear, deviating from the confidence pattern on the previous task.

Network/ n-gram		Mean Luce ratio difference	Mean Luce ratio hits	Mean Luce ratio misses
srn	5hu	0.4346 ± 0.2551	0.3961 ± 0.0977	0.3853 ± 0.0971
	10hu	0.4549 ± 0.2434	0.4007 ± 0.0399	0.4486 ± 0.1023
	25hu	0.4036 ± 0.3033	0.4349 ± 0.0456	0.3644 ± 0.0325
rbp	5hu	0.4418 ± 0.2161	0.6440 ± 0.1105	0.5557 ± 0.1194
	10hu	0.4701 ± 0.1707	0.6273 ± 0.1550	0.6743 ± 0.1507
	25hu	0.3805 ± 0.2634	0.6642 ± 0.1053	0.6093 ± 0.0973
2gram		0.3696 ± 0.2785	0.4832 ± 0.0821	0.4488 ± 0.0812
4gram		0.0690 ± 0.1448	0.8040 ± 0.1882	0.4804 ± 0.1423
6gram		0.0316 ± 0.0586	0.8472 ± 0.2011	0.4470 ± 0.0908
8gram		0.0478 ± 0.0921	0.8410 ± 0.2116	0.4703 ± 0.1158

Table 3.12. Confidence in predictions concerning identity recursive structures.

Embedding Depth Performance

Looking at Figure 3.12, we can see that network performance across embedding depth follows a pattern comparable with the one found on mirror recursion; viz., performance degrades as a function of embedding depth (except for the RBP network with 5 hidden

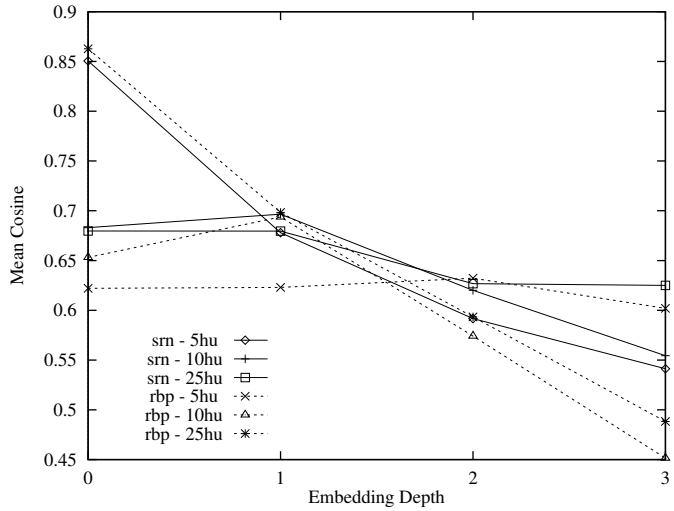


Figure 3.12. Network performance regarding identity recursion plotted against embedding depth.

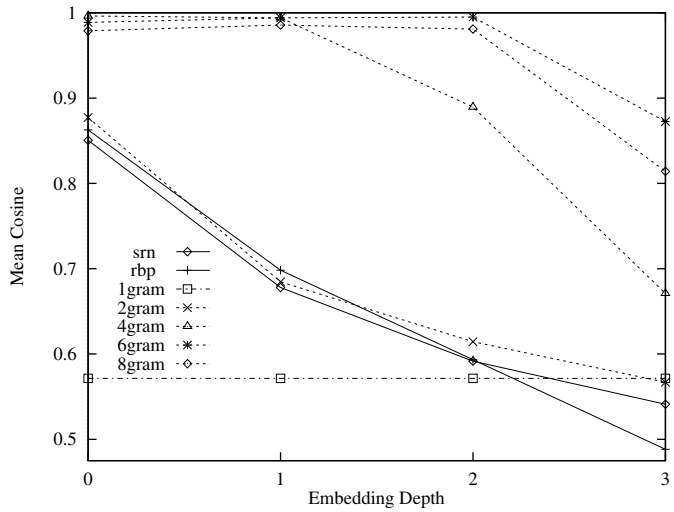


Figure 3.13. Network (SRN: 5hu; RBP: 25hu) and n-gram performance regarding identity recursion plotted against embedding depth.

units). This was more or less what was predicted in Section 3.1.1. Figure 3.13 illustrates that bigram based predictions follow the same trend. However, prediction based on higher n-grams (especially hexagrams and octograms) are performing remarkably well—considerably better than on the mirror recursion task. This might be explained by the fact that the distance between the first noun and its verb is shorter in an deeply embedded identity recursive sentence compared with a mirror recursive sentence of the same embedding depth. In addition, both network and n-gram based performance is always above the level of performance that can be obtained by making predictions according to the simple relative frequency of the words in the data set.

To recapitulate, the nets did better than expected on this task–even though their performance is slightly worse than on the mirror recursion task. Once again, performance degraded across depth of embedding which is also the case for human performance on similar (cross-dependency) structures. However, the higher n-gram predictions did not display the same degraded performance, suggesting that they are not suitable as a basis for a human performance model. And once more, net performance was very close to that exhibited by bigram based predictions.

3.3.4 Summary

The performance of the two nets on the three tasks turned out to be close to the performance predictions made in Section 3.1.1. However, a few things are worth noting. First of all, the overall performance (of both nets and n-gram based predictions) on the identity recursion task was somewhat better than expected. This is a positive result, given that dealing with identity recursive structures require the acquisition of (something closely related to) a context-sensitive grammar. Secondly, there was no significant performance difference between the two kind of networks on either of the tasks (albeit that the RBP network generally was more confident in its predictions). Thirdly, network performance was on the same level as performance obtained on bigram based predictions. And finally, the size of the hidden layer did not seem to influence performance—in particular, bigger nets did not do better (although there is a weak tendency towards better confidence for bigger nets).

3.4 Experiment 2: Eight Word Vocabulary

In order to test further the ability of both networks to capture the recursive regularities necessary for dealing with novel sentences, I conducted a second experiment involving

an eight word vocabulary[15]. Thus, we have four nouns in a singular (*'a', 'b', 'c', 'd'*) and a plural form (*'A', 'B', 'C', 'D'*), an four verbs likewise in a singular (*'w', 'x', 'y', 'z'*) and a plural form (*'W', 'X', 'Y', 'Z'*). In experiment 1, I found that the size of the hidden unit layer did not appear to influence performance on either of the tasks. I therefore decided only to train networks with 20 hidden units in the present experiment. The SRN was trained for 400 epochs and the RBP net for 200 epochs (using the same back-propagation parameters as in experiment 1). Throughout experiment 2 the training sets contained 2000 sentences and the test sets 1000 sentences.

Pilot studies indicated that the localist representation of words that I used in the previous experiment was inappropriate for the current experiment. Instead, I adopted a different representation scheme in which each word was represented by a single bit (independently of its number) and the number was represented by one of two bits (common to all words) signifying whether a word was singular or plural. Thus, for each occurrence of a word two bits would be on—one bit signifying the word and one bit indicating its number[16]. The input/output consisted of 11 bit vectors (one for each of the eight words, one for each of the two word numbers (singular or plural), and one for the end of sentence marker). To allow assessment of network performance on novel sentences, I introduced two extra test sets with, respectively, 10 novel sentences and 10 previously seen sentences (all with mean length: 5.30; sd: $\pm$1.64).

3.4.1 Counting Recursion

In this version of the counting recursion task, the nets were trained on a training set consisting of counting recursive sentences of variable length (mean: 4.73; sd $\pm$1.33) and tested on a separate large data set (mean: 4.72; sd: $\pm$1.34) as well as on two small test sets consisting of, respectively, novel and previously seen sentences. The embedding distribution of the two large data sets is shown in table 3.13.

[15]This extension of the vocabulary was necessary, since leaving out certain sentence structures in the previous experiment would have skewed the training set in a problematic fashion. Moreover, I wanted to investigate how the networks would perform with a bigger vocabulary.

[16]It is worth noticing that this kind of representational format appears more plausible than a strict localist one. In particular, it is unlikely that we 'store' singular and plural forms of the same word (e.g., 'cat' and 'cats') as distinct and completely unrelated representations as it would be the case with localist representations. Rather, I would expect the human language processing mechanism to take advantage of the similarities between the two word forms to facilitate processing. Derivational morphology could, perhaps, be construed as the instantiation of such a system

Embedding	Depth 0	Depth 1	Depth 2	Depth 3	Depth 4
Training set	29.20%	55.70%	14.45%	0.65%	0.00%
Test set	29.50%	55.90%	13.80%	0.70%	0.10%

Table 3.13. The distribution of different depths of embedding in the training and test sets involving counting recursion.

General Performance

Both nets performed slightly worse on this version of the counting recursion task compared with the previous version. As can be seen from table 3.14, the SRN is doing better than the RBP net. The latter is not doing better than performance based on the relative frequency of the words in the training set (1gram), perhaps suggesting that the RBP net has not quite picked up the sequential structure of the data. Still, comparisons between performance before and after training (presented in table 3.15 and 3.14, respectively) indicates that performance improved considerably through learning. Thus, the SRN increased its performance 30% (from $cos = 0.603$ to $cos = 0.789$) and the RBP net 58% (from $cos = 0.466$ to $cos = 0.737$).

Network/ n-gram	Mean squared error	Mean Cosine
srn	0.3972 ± 0.4146	0.7894 ± 0.1722
rbp	0.4870 ± 0.5412	0.7371 ± 0.2324
1gram	0.4487 ± 0.4269	0.7327 ± 0.2336
2gram	0.3004 ± 0.4316	0.8424 ± 0.1903
4gram	0.1867 ± 0.4814	0.9218 ± 0.1746
6gram	0.5342 ± 0.6257	0.7198 ± 0.3188
8gram	0.9367 ± 0.4230	0.1898 ± 0.3167

Table 3.14. General performance on counting recursive structures.

Network	Mean squared error	Mean Cosine
srn	1.3277 ± 0.4506	0.6029 ± 0.1451
rbp	3.4272 ± 0.8095	0.4657 ± 0.1928

Table 3.15. Baselines for the general performance on counting recursive structures tested on the full conditional probabilities of the test set.

However, it should be noted that the high level of performance obtained by the simple relative frequency predictions is due to the semi-arbitrariness of the counting recursion task. As there are no agreement constraints between nouns and verbs in this

task, predicting the next word according to the relative frequency of the words in the training set will result in decent performance. In this experiment the bigram based performance is slightly better than the SRN performance. Yet, the biggest difference is that the higher n-gram based performance is quite poor (save quadrogram based predictions). As in experiment 1, this is because there are too many single occurrences of unique higher n-grams in the training set. This impairs generalization from the training set to the test set.

Network	Novel sentences (mean cosine)	Previously seen sentences (mean cosine)
srn	0.4594 ± 0.2583	0.4819 ± 0.2404
rbp	0.4050 ± 0.2603	0.3909 ± 0.2747

Table 3.16. Network performance on novel and previously seen sentences involving counting recursive structures.

Table 3.16 shows net performance on the novel and previously seen sentences. It is clear that both networks have acquired the ability to generalize to novel sentences[17]. There is only a small decrease in performance when comparison is made between the two types of test sentences. The performance of the SRN degrades by less than 5%, whereas the performance of the RBP net actually improves very slightly by 4%. It is likely that the RBP network performed slightly better on the novel sentences compared with previously seen sentences because the network might have been somewhat undertrained.

Embedding Depth Performance

Turning to performance across embedding depth, Figure 3.14 shows that the general trend from experiment 1 is replicated in the present task: performance degrades as a function of embedding depth. Both nets exhibit the same behavior across embedding depth—though the RBP net performance is slightly poorer than that of the SRN. The bigram performance is closer to the SRN performance than table 3.14 suggests. Quadrogram based predictions do well at the first two embedding depths, but degrades rapidly after that. Hexagram and octogram based performance are doing very poorly—especially the latter (for the reasons mentioned above). Most strikingly, predictions

[17] Noe that this apparently low performance is due to the fact that it was measured against the probability distribution of these two sets, whereas the nets had been trained on (and, thus, become sensitive to) the much more complex probability distribution of the 2000 sentences in the training set. In addition, the embedding distribution in these test sets were skewed slightly towards longer sentences (hence the higher mean sentence length reported above).

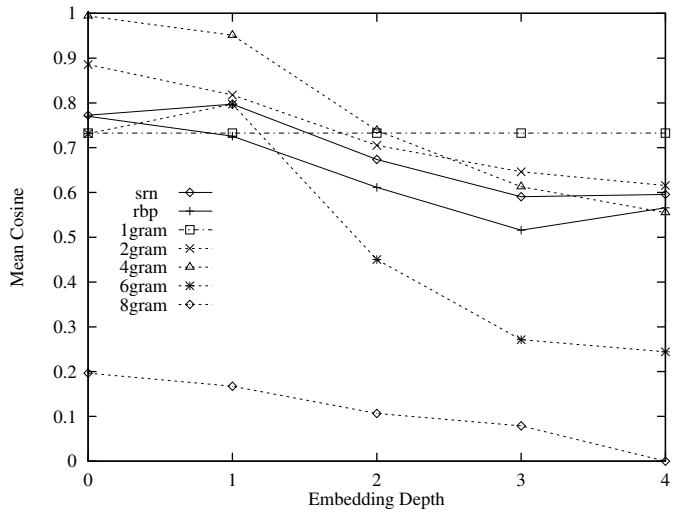


Figure 3.14. Network and n-gram performance regarding counting recursion plotted against embedding depth.

relying entirely on the relative frequency of the words in the training set are superior to all other kinds of prediction methods from depth 2 onwards.

To sum up, both nets did slightly worse on this task than on the same task in the previous experiment. Nevertheless, the nets were able to generalize to novel sentences. On the other hand, predictions made from the simple relative word frequency did surprisingly well, suggesting that the semi-random nature of the counting recursion task makes it difficult to learn for larger vocabularies.

3.4.2 Mirror Recursion

Both networks were trained on exclusively mirror recursive sentences (mean: 4.77; sd: $\pm$1.36) and tested on a large separate data set (mean: 4.68; sd: $\pm$1.35) as well as two 10 sentence data sets with, respectively, novel and previously seen sentences. Table 3.17 shows the embedding distribution in the two large data sets.

Embedding	Depth 0	Depth 1	Depth 2	Depth 3	Depth 4
Training set	28.50%	55.65%	14.85%	0.95%	0.05%
Test set	31.10%	54.30%	13.90%	0.70%	0.00%

Table 3.17. The distribution of different depths of embedding in the training and test sets involving mirror recursion.

General Performance

On this task the SRN, once more, performed modestly better than the RBP network as can be seen from table 3.18. In contrast to the previous task, both nets achieve a level of performance that is better than what can be accomplished by making predictions based on simple relative word frequency (1gram). Moreover, the performance on this task

Network/ n-gram	Mean squared error	Mean Cosine
srn	0.3499 ± 0.4202	0.8127 ± 0.1931
rbp	0.4838 ± 0.4901	0.7756 ± 0.1979
1gram	0.4492 ± 0.4304	0.7328 ± 0.2361
2gram	0.2984 ± 0.4314	0.8443 ± 0.1888
4gram	0.1985 ± 0.4944	0.9163 ± 0.1786
6gram	0.5393 ± 0.6404	0.7174 ± 0.3209
8gram	0.9473 ± 0.4132	0.1862 ± 0.3113

Table 3.18. General performance on mirror recursive structures.

(though somewhat poorer than on the counting recursion task) is comparable with the performance on the same task in experiment 1. The n-gram based performance is very close to that obtained on the previous task. Bigram based predictions are still slightly better than net predictions. A comparison between the performance of the untrained (table 3.19) and the trained (table 3.18) nets reveals that the RBP net had a much higher relative performance improvement through learning (67% – from $cos = 0.464$ to $cos = 0.776$) than the SRN (35% – from $cos = 0.602$ to $cos = 0.813$).

Network	Mean squared error	Mean Cosine
srn	1.3313 ± 0.4512	0.6016 ± 0.1451
rbp	3.4352 ± 0.8087	0.4635 ± 0.1926

Table 3.19. Baselines for the general performance on mirror recursive structures.

From table 3.20 it can be seen that the networks exhibited no significant difference in performance on, respectively, the novel and the previously seen test sentences. Thus, the nets have learned to generalize to novel sentences. The performance of the SRN is practically the same on both kinds of sentences (a difference of less than 1%). The RBP net, again, had an increase in performance (9%) on the novel test sentences compared with the sentences it had already been exposed to during training (once more suggesting, perhaps, undertraining). Notice also that the SRN is doing better on both kinds of sentences compared with the previous task. In the same way, the RBP

net performs better on the novel sentences on this task.

Network	Novel sentences (mean cosine)	Previously seen sentences (mean cosine)
srn	0.5546 ± 0.2007	0.5504 ± 0.1995
rbp	0.4370 ± 0.2703	0.3994 ± 0.2846

Table 3.20. Network performance on novel and previously seen sentences involving mirror recursive structures.

Embedding Depth Performance

Figure 3.15 shows that performance as a function of embedding depth exhibits much the same general pattern of degradation as found on the same task in the previous experiment (except from a minor peak at depth 1). Once again, we see that the performance of the nets is comparable with that of bigram predictions. As in the counting recursion task, the prediction based on quadrograms obtain the best performance—though it degrades to the level of the nets at depth 2—whereas hexagram and octogram based predictions do badly. Compared with the previous task, the nets accomplish a better performance in relation to performance based on simple relative word frequency predictions (1gram); though the latter is still the superior prediction method for depths 2 and 3.

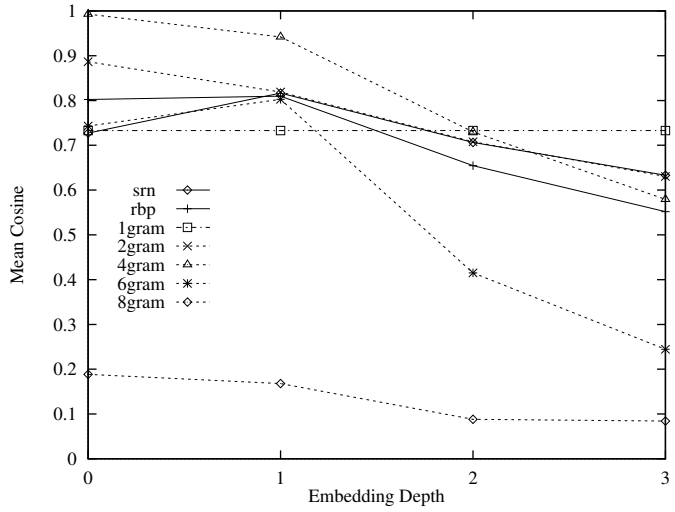


Figure 3.15. Network and n-gram performance regarding mirror recursion plotted against embedding depth.

In summary, the nets performed satisfactorily on this task—especially on novel

sentences. The SRN reached the same level of performance as on the mirror recursion task in the previous experiment. The RBP net performed slightly worse.

3.4.3 Identity Recursion

On the final task, the networks were trained on a identity recursion training set (mean: 4.75; sd: $\pm$1.34) and tested on a similar, but separate, data set (mean: 4.70; sd: $\pm$1.33) as well as two 10 sentence test sets comprised of, respectively, novel and previously seen sentences. Table 3.21 presents the distribution of embeddings in the two large data sets.

Embedding	Depth 0	Depth 1	Depth 2	Depth 3
Training set	28.70%	55.90%	14.55%	0.85%
Test set	29.70%	54.10%	13.40%	0.80%

Table 3.21. The distribution of different depths of embedding in the training and test sets involving identity recursion.

General Performance

The overall performance of the two nets was much alike, though favoring the SRN. This is in contrast to the relative increase in performance achieved through learning, where the RBP network obtained a 58% improvement (from $cos = 0.476$ to $cos = 0.755$) compared with the SRN's 28% (from $cos = 0.604$ to $cos = 0.773$) (cf. table 3.22 and 3.23). The performance of both nets on identity recursion was better in this

Network/ n-gram	Mean squared error	Mean Cosine
srn	0.4097 $\pm$ 0.4417	0.7732 $\pm$ 0.1760
rbp	0.4629 $\pm$ 0.5070	0.7546 $\pm$ 0.2002
1gram	0.4456 $\pm$ 0.4217	0.7326 $\pm$ 0.2325
2gram	0.2962 $\pm$ 0.4214	0.8429 $\pm$ 0.1866
4gram	0.1918 $\pm$ 0.4773	0.9174 $\pm$ 0.1739
6gram	0.5371 $\pm$ 0.6202	0.7129 $\pm$ 0.3253
8gram	0.9375 $\pm$ 0.4147	0.1849 $\pm$ 0.3127

Table 3.22. General performance on identity recursive structures.

experiment compared with the same task in experiment 1 (but worse than on the previous two tasks—which is in accordance with the performance predictions made in Section 3.1.1). All the n-gram based performances closely mirrored the performance on the previous two tasks.

Network	Mean squared error	Mean Cosine
srn	1.3229 ± 0.4466	0.6036 ± 0.1442
rbp	3.2900 ± 0.7834	0.4763 ± 0.1903

Table 3.23. Baselines for the general performance on identity recursive structures.

Most importantly, as it was the the case in the previous tasks, both networks were able to deal with novel sentences, indicating that they had learned the underlying recursive regularities. Table 3.24 contains the results from the testing of the two nets on novel and previously seen sentences. The SRN performance on novel sentences degraded by only 2% compared with its performance on previously seen sentences. Once again, the RBP net achieved a small performance improvement of 3% on novel sentences compared with its performance on sentences that had been presented to it during training.

Network	Novel sentences (mean cosine)	Previously seen sentences (mean cosine)
srn	0.5376 ± 0.2436	0.5485 ± 0.2451
rbp	0.4762 ± 0.2473	0.4629 ± 0.2828

Table 3.24. Network performance on novel and previously seen sentences involving identity recursive structures.

Embedding Depth Performance

Figure 3.16 illustrates the close fit between the performance of the two networks across embedding depth. It also shows that the nets are not as close to the bigram performance as in the previous task (and in the previous experiment). As in the previous two tasks, both hexagram and octogram based predictions reach a poor level of performance (especially, the latter). Moreover, the performance of both nets is again inferior to the quadrogram based performance. For embedding depths 2 and 3, predictions relying on the simple relative frequency of the words in the training set are still superior to all other prediction methods.

In short, both nets also did well on this version of the identity recursion task. In particular, they were able to generalize what they had learned to novel sentences—even though both nets performed slightly worse on this task compared with the previous two tasks in the current experiment.

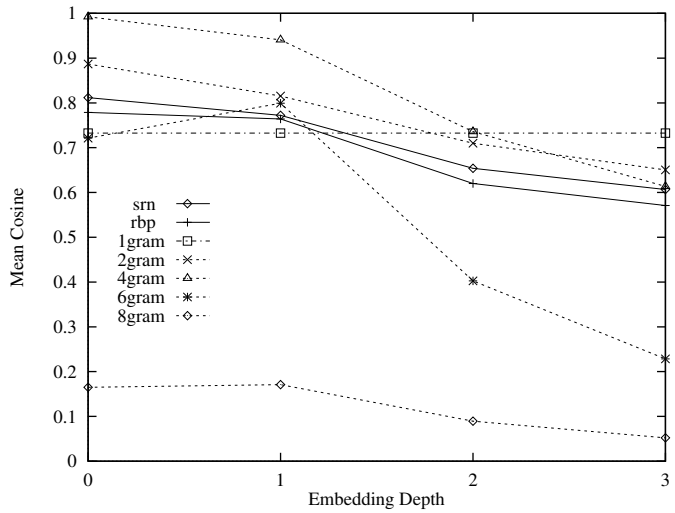


Figure 3.16. Network and n-gram performance regarding identity recursion plotted against embedding depth.

3.4.4 Summary

The two nets performed in much the same way in the present experiment as in the previous experiment. Thus, the performance predictions outlined in Section 3.1.1 were confirmed once more (again with the exception of the relatively good performance achieved in the identity recursion task). However, there are a few differences between the two experiments. First, in the current experiment the SRN achieved better results on all three tasks compared with the RBP net—though there are some indications that the latter might have been undertrained. Secondly, both nets accomplished a satisfactory level of generalization when faced with novel sentences. This suggests that the nets were able to learn the recursive regularities underlying the three tasks. Thirdly, for all three tasks predictions based on the relative word frequencies of the training set reached a surprisingly high level of performance. This is presumably due to the new representation format. Finally, the better performance on novel sentences on the mirror and identity recursion tasks (measured in terms of mean cosines) compared with the counting recursive task indicates that the learning of generalization relevant information is facilitated by the agreement constraints.

3.5 Discussion

In this chapter I have addressed Chomsky's (1957) challenge in a slightly reformulated form: Can neural networks capture the ni-recursive regularities of natural language

if we accept that arbitrarily complex sentences cannot (and, perhaps, should not) be handled? The ability of both kinds of networks to generalize to novel sentences involving three kinds of ni-recursion (in experiment 2) suggests that neural networks might be able to do the trick. But where does that leave the pattern of gradual breakdown of performance as observed in all the simulations presented here? If we compare the breakdown pattern in the mirror and identity recursion tasks with the degradation of human performance on center-embedded and cross-dependency structures (as can be adduced from Figure 3.17[18]), we can conclude that such a breakdown pattern is, indeed, desirable from a psycholinguistic perspective. Thus, network (and bigram based) performance across embedding depth appears to mirror general human limitations on the processing of complex ni-recursive structures. Moreover, given the performance on the counting recursion task we can make the empirical prediction that human behavior on nested *'if–then'* structures will have the same breakdown pattern as observed in relation to the nested center-embedded and cross-dependency sentences (though with a slightly better overall performance). That is, I predict that humans are only able to process a very limited nesting of *'if–then'* structures, and that the performance, furthermore, will exhibit graceful degradation over depth of nesting.

Two other things are worth noting. First of all, the overall performance (of both nets and n-gram based predictions) on the identity recursion task was better than expected. This is a positive result, given that dealing with identity recursive structures requires the acquisition of (something closely related to) a context-sensitive grammar, whereas mirror recursion 'merely' requires the acquisition of a context-free grammar. The networks, then, did better on the identity task than was to be expected given the structural complexity of the learning task (as outlined in section 3.1.1). This is important, since human performance seems to be quite similar on both kinds of ni-recursive structure (see Figure 3.17). Secondly, there was no significant performance difference between the two kinds of networks on either of the tasks (in both experiments). This means that the negative results reported by Chater & Conkey (1992) regarding SRN performance on certain non-language tasks do not extend themselves to more language-like tasks. Thus, in addressing my secondary motivation for the present simulations, we found, rather surprisingly, that unfolding a recurrent network for the purpose of RBP does not seem to provide additional computational power on the language-like tasks presented here.

[18]The data from Bach, Brown & Marslen-Wilson (1986: p. 255, table 1: test results) is displayed by reversing the scale so that it is readily comparable with the simulation results expressed in terms of mean cosines. This amounts to plotting the y-coordinates as $f(n) = 9 - n$, where n is the original data point.

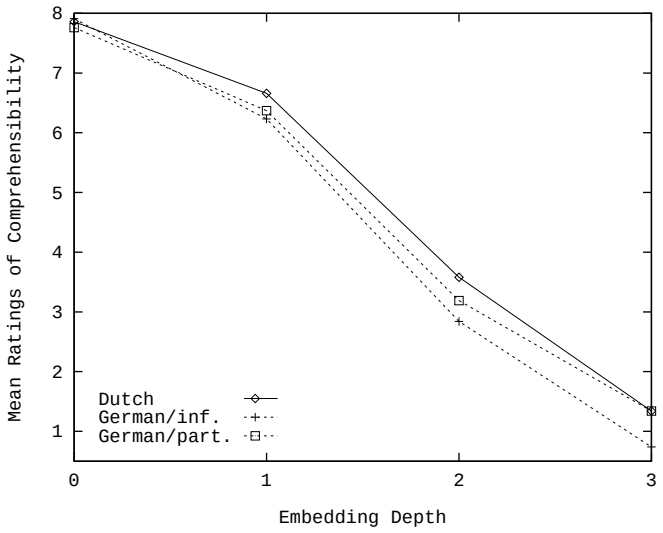


Figure 3.17. The performance of native speakers of German and Dutch on, respectively, center-embedded (mirror recursive) sentences and sentences involving cross-dependencies (identity recursion) is plotted against embedding depth. The figure is based on data reported in Bach, Brown, & Marslen-Wilson (1986).

The close similarity between the breakdown patterns in human and neural network performance on complex ni-recursive structures supports two wide-reaching conjectures. On the one hand, neural network models—in spite of their finite-state nature—must be considered as viable models of natural language processing. At least, I have shown that the existence of center-embedding and cross-dependency no longer can be used as *a priori* evidence against neural network (and other finite state) models of linguistic behavior. On the other hand, the common pattern of graceful degradation also suggests that humans, like neural networks, are sensitive to the statistical structure of language. Neural networks pick up certain simple statistical contingencies in the input they receive (the simulations presented here indicate that such statistics might resemble bigram based probability distributions[19]). I suggest that the breakdown pattern in human performance on complex recursive structures might also be due to a strong dependence on such statistics in the acquisition of linguistic structure. Whether these conjectures are true is a matter of future empirical research, not *a priori* speculation.

This chapter has focused on addressing the Chomskyan (1957) challenge expressed in terms of his three abstract languages involving ni-recursion. The results presented

[19]However, it should be noted that the very good n-gram results presented in this chapter should be taken with some caution. The next chapter shows that an SRN attains better performance than n-gram based predictions when faced with a more realistic grammar incorporating both i- and ni-recursion.

here are obviously somewhat limited by the fact that only one kind of recursive structure occurs in each of the three languages. In the next chapter, I therefore present simulations replicating the results found here, but in the context of a grammar incorporating not only ni-recursion but also different instances of i-recursion.

Chapter 4

Connectionist Learning of Linguistic Structure

It is often noted that natural language recursion poses a serious problem for connectionist and other finite state models of language. Indeed, the existence of recursion in any cognitive domain appears to be highly problematic for non-symbolic approaches to cognition because recursion *qua* computational property is defined in essentially symbolic terms (Fodor & Pylyshyn, 1988). This problem becomes even more pressing since most classical models of cognition rely on recursion to achieve infinite productivity. For example, the *'language of thought'* hypothesis (Fodor, 1975) involves the processing of recursive structures with respect to most—if not all—kinds of cognitive behavior. However, in their recent defense of the classical position against the threat of connectionism, Fodor & Pylyshyn (1988) concentrate on language as a paradigm case for symbolic processing involving recursion. More specifically, they argue that "the empirical arguments for [recursive] productivity have been made most frequently in connection with linguistic competence" (p. 34); and that "linguistic capacity is a paradigm of systematic cognition" (p. 37). Consequently, the productivity and systematicity arguments they provide in favor of the classical position are all based either directly on language, i.e., linguistic capacity, or indirectly on language related behavior, i.e., verbal reasoning.

Crucially, the existence of recursion in natural language presupposes that the grammars of linguistic theory correspond to *real* mental structures, rather than mere structural *descriptions* of language *per se*. Yet, there are no *a priori* reasons for assuming that the structure of the observable public language necessarily must dictate the form of our internal representations (Stabler, 1983; van Gelder, 1990b). Nevertheless, whatever system of internal representations we might want to posit instead of the traditional

linguistic grammars, it must be able to account for the diversity of language behavior. As pointed out in chapter 2, connectionist models are able to deal with a certain kind of recursion: iterative recursion (i-recursion). Chapter 3 demonstrated that recurrent neural networks moreover are capable of processing a limited amount of non-iterative recursion (ni-recursion) and, furthermore, exhibit a graceful degradation of performance comparable with that of humans. However, it remains to be seen whether such models can learn the regularities underlying grammars involving *both* i- and ni-recursion, and, at the same time, display a human-like performance.

In the present chapter, I address this question via two simulation experiments in which simple recurrent networks are trained on grammars incorporating left- and right i-recursion as well as ni-recursion, either in the form of center-embeddings or cross-dependencies. The first section describes the two grammars used in the experiments. The training regime adopted here is that of incremental memory learning (Elman, 1991b, 1993) which simulates the effects of maturational constraints on the acquisition process. The next section reports the results, with general performance first followed by the performance on ni- and i-recursive structures, respectively. Section 3 discusses the issue of generalization in connectionist models, outlines a formalization of linguistic generalization, and presents simulation results suggesting that connectionist models may be able to attain a powerful kind of generalization. Finally, I discuss the prospects for this type of networks as models of language learning and processing.

4.1 Learning Complex Grammars

The simulations presented in this chapter build on, and extend, the results from chapter 3 as well as simulation work reported in Chater (1989), Chater & Conkey (1992), Cleeremans, Servan-Schreiber & McClelland (1989), Servan-Schreiber, Cleeremans & McClelland (1989, 1991), and most notably in Christiansen & Chater (1994), Elman (1991a, 1991b , 1992, 1993), and Weckerly & Elman (1992). In the present simulation experiments, a simple recurrent network was trained to derive grammatical categories given sentences generated by a phrase structure grammar. Two grammars were used in these experiments, both involving left- and right i-recursion but differing in the kind of ni-recursion they incorporated. Whereas the grammar shown in Figure 4.1 allows center-embedding, the one illustrated in Figure 4.2 affords cross-dependencies. Both grammars use the same small vocabulary, presented in Figure 4.3, consisting of two proper nouns, three singular nouns, five plural nouns, eight verbs in both plural and singular form, a singular and a plural genitive marker, three prepositions, and three ('locative') nouns to be used with the prepositions.

S → NP VP "."
NP → PropN | N | N rel | N PP | gen N | N *and* NP
VP → V(i) | V(t) NP | V(o) (NP) | V(c) *that* S
rel → *who* NP V(t|o) | *who* VP
PP → prep prepN
gen → N + "s" | gen N + "s"

Figure 4.1. The phrase structure grammar incorporating center-embedded as well as left- and right-recursive structures.

S → NP VP "."
S → N_1 N_2 $V(t|o)_1$ $V(i)_2$ | N_1 N_2 N $V(t|o)_1$ $V(t|o)_2$
S → N_1 N_2 N_3 $V(t|o)_1$ $V(t|o)_2$ $V(i)_3$ | N_1 N_2 N_3 N $V(t|o)_1$ $V(t|o)_2$ $V(t|o)_3$
NP → PropN | N | N rel | N PP | gen N | N *and* NP
VP → V(i) | V(t) NP | V(o) (NP) | V(c) *that* S
rel → *who* VP
PP → prep prepN
gen → N + "s" | gen N + "s"

Figure 4.2. The phrase structure grammar incorporating crossed dependencies as well as left- and right-recursive structures.

PropN → {*John, Mary*}
N → {*boy, boys, girl, girls, man, men, cats, dogs*}
V(i) → {*jumps, jump, runs, run*}
V(t) → {*loves, love, chases, chase*}
V(o) → {*sees, see*}
V(c) → {*thinks, think, says, say, knows, know*}
prep → {*near, from, in*}
prepN → {*town, lake, city*}

Figure 4.3. The vocabulary used with the two grammars (singular forms are placed before their corresponding plural forms where appropriate).

Importantly, the grammars are significantly more complex than the one used by Elman (1991a, 1992). The latter involved subject noun/verb number agreement, verbs which differed with respect to their argument structure (intransitive, transitive, and optionally transitive verbs), and relative clauses (allowing for multiple embeddings with complex agreement structures). I have extended this grammar by adding prepositional modifications of noun phrases (e.g., *'boy from town'*), left recursive genitives (e.g., *'Mary's boy's cats'*), conjunction of noun phrases (e.g., *'John and Mary'*), and sentential complements (e.g., *'John says that Mary runs'*). In the cross-dependency grammar, the object relative clause (which creates center-embedding) has been removed. Instead, four additional expansions of S have been added to allow for crossed dependencies (see section 4.2.2 for further explanation)[1]. Both grammars can generate the following sample sentences involving i-recursive structures:

Mary knows that John's boys' cats see dogs.
boy loves girl from city near lake.
man who chases girls in town thinks that Mary jumps.
John says that cats and dogs run.
Mary who loves John thinks that men say that girls chase boys.

In addition, the *center-embedding* grammar are able to produce ni-recursive sentences such as:

girl who men chase loves cats.
cats who John who dogs love chases run.

These two sentences can be rephrased in terms of subject relative clauses; that is, as *'men chase girl who loves cats'* and *'dogs love John who chases cats who run'*, respectively. The *cross-dependency grammar*, on the other hand, can express the same sentential content in this way:

men girl cats chase loves.
dogs John cats love chases run.

Notice that the cross-dependency grammar can also rephrase these two sentences in terms of subject relative clauses.

[1]The cross-dependency grammar is supposed to correspond to a Dutch grammar, even though the vocabulary used for convenience is English (Figure 3). This also means that the semantic constraints on cross-dependency structures in Dutch are likely to be violated. As pointed out to me by Pauline Kleingeld (a native Dutch speaker), Dutch cross-dependency structures are limited to constructs expressing something which can be observed together. However, the lack of such semantic constraints are not important for the present simulations, since the latter only deals with syntax.

The results from the previous chapter showed no marked difference in the learning ability of simple recurrent networks (Elman, 1988, 1989, 1990, 1991a) compared with networks using back-propagation through time (Rumelhart, Hinton & Williams, 1986). I therefore chose to use only the former in the present simulations because it is much less expensive in computational terms, and, perhaps, more cognitively plausible. Recall that this kind of network is a standard feedforward network equipped with an extra layer of so-called context units to which the activation of the hidden unit layer at time t is copied over and used as additional input at $t + 1$ (see also Figure 3.5 in chapter 3). The nets used here had 42 input/output units[2] and 150 hidden units (a 42–150–42 configuration) with an additional 150 context units. Each net was trained on a next word prediction task using incremental memory learning as proposed by Elman (1991b, 1993), providing the net with a memory window which "grows" as training progresses. Pilot simulations had suggested using a learning rate of 0.1, no momentum, and an initial randomization of the weights between ± 0.001. All training sets used in the simulations consisted of 10,000 randomly generated sentences of variable length and complexity. Sentences in the training sets generated by the center-embedding grammar had an approximate mean length of 6 words (sd: ± 2.5). The cross-dependency grammar also produced sentences with a mean length of about 6 words (sd: ± 2.0).

The training, using the incremental memory learning strategy, progressed as follows: First, the center-embedding net was trained for 12 epochs and the cross-dependency net for 10 epochs on their respective training sets. To simulate an initially limited memory capacity, the context units were reset randomly after every three or four words. The training sets were then discarded, and both nets trained for 5 epochs on new training sets, now with a memory window of 4–5 words. This process was repeated for two consecutive periods of 5 epochs, each on different training sets with a memory window of 5–6 words and 6–7 words, respectively. Finally, the nets were trained for 5 epochs on new training sets, this time without any memory limitations[3]. The growing memory window is assumed to reflect decreasing constraints on a child's memory abilities following maturational changes, ending up with the full adult system (Elman, 1993). Other simulations (not reported in detail here) have shown that 'adult' networks—that

[2]Of the 42 input/output units only 38 were used to represent the vocabulary and other lexical material in a localist format (the remaining four units were saved for other purposes not mentioned further here). The choice of the localist representation over the compressed representation used in experiment 2 in chapter 3 was motivated by the need to separate simultaneous activations of both single and plural items.

[3]Although the center-embedding net was trained for 32 epochs and the cross-dependency net for 30 epochs, preliminary results from other simulations, presently underway, suggest that optimal performance is reachable after much less training.

is, nets not undergoing simulated maturational changes—cannot learn the grammatical regularities of the two grammars in a satisfactory way[4]. This corroborates similar results presented in Elman (1991b, 1993) and in Goldowsky & Newport (1993), suggesting that the initial memory limitations of children may, at least in part, help with the bootstrapping of linguistic structure. Data from first and second language learning further supports this *'less is more'* hypothesis (Newport, 1990—for further discussion of maturational constraints on learning, see chapter 5, section 5.3)

4.2 Results

In order to provide an independent basis for the assessment of general network performance on the two grammars, the simple statistical prediction method developed in connection with the previous chapter, was adjusted to produce n-gram given the data from either grammar. It should be noted that this program was only trained on the training set that the networks saw in their final five epochs of training. However, given the size of the training sets this should not lead to a significant decrease in performance compared with that of the nets.

4.2.1 General Performance

Both the net trained on sentences from the center-embedding grammar and the net trained on sentences from the cross-dependency grammar performed very well. Their overall performance was comparable with general net performance in the two experiments reported in chapter 3 (see also section 4.3.4 on generalization tests). More interestingly, both nets were able to surpass n-gram based performance, indicating that the nets are doing more than just learning n-gram statistics.

Performance on the Center-embedding Grammar

The general performance on the center-embedded grammar in Figure 4.1 was assessed on a test set consisting of 10,000 randomly generated sentences (mean length: 6.13; sd: $\pm$2.53). The results are presented in table 4.1[5]. The trained network is doing slightly better than n-gram based prediction in terms of mean cosines, but reaching the same level of mean squared error performance as trigram based prediction. Looking at the

[4]The adult center-embedding and cross-dependency nets were trained for 32 and 30 epochs, respectively.

[5]As in chapter 3, the results were measured not against the target, but against the full conditional probabilities in order to take the indeterministic nature of the prediction task into account (see also, chapter 3, footnote 8).

Network/ n-gram	Mean squared error	Mean Cosine
trained net	0.2169 ± 0.3523	0.7904 ± 0.2702
untrained net	9.9506 ± 0.4075	0.3264 ± 0.1527
1gram	0.4093 ± 0.3650	0.3851 ± 0.2181
2gram	0.2662 ± 0.3668	0.7084 ± 0.2909
3gram	0.2255 ± 0.3629	0.7843 ± 0.2948
4gram	0.2636 ± 0.4142	0.7335 ± 0.3118
5gram	0.4225 ± 0.5222	0.5499 ± 0.3579

Table 4.1. General performance of the net trained on the center-embedding grammar.

standard deviation we can also see that the net is more consistent in its predictions (albeit all standard deviations are relatively high[6]). Notice the 97% decrease of the mean squared error (from 9.9506 for the untrained net to 0.2169 for the trained net). The training induced increase in performance measured via the mean cosine is an impressive 142%—almost 2.5 times better than the untrained performance (from $cos = 0.3264$ to $cos = 0.7904$). Together with the low performance obtained by merely making predictions according to the basic relative frequency of the words in the training set (1gram) this strongly suggests that a considerable amount of sequential structure has been learned by the net. Moreover, although the overall performance of the adult net (MSE: 0.2142 ± 0.3467; mean cosine: 0.7989 ± 0.2698) is comparable with the that of the maturationally constrained net, a more detailed analysis of the former's predictions showed that it performed significantly worse on complex structures such as center-embedding.

Performance on the Cross-dependency Grammar

The network trained on sentences produced by the cross-dependency grammar in Figure 2 was tested for general performance on 10,000 randomly generated sentences (mean length: 5.98; sd: ± 2.01), and the results reported in table 4.2. This net was able to surpass n-gram based predictions both in terms of mean cosine and mean squared error, distancing itself more clearly from trigram based performance than in the center-embedding case above. Overall, standard deviations are similar to what was found in the previous simulation (that is, still rather high). Again, we see a significant increase in performance as a consequence of learning. The mean squared error was reduced by

[6]These high standard deviations hide the fact that the errors made by both the center-embedding and the cross-dependency nets generally increased gradually with sentence length (see the detailed results below), whereas n-gram errors remained slightly more constant.

Network/ n-gram	Mean squared error	Mean Cosine
trained net	0.1931 ± 0.3436	0.8100 ± 0.2516
untrained net	9.9283 ± 0.4162	0.3597 ± 0.1989
1gram	0.4014 ± 0.3722	0.4304 ± 0.2155
2gram	0.2758 ± 0.3697	0.6547 ± 0.3116
3gram	0.2264 ± 0.3655	0.7646 ± 0.3014
4gram	0.2618 ± 0.4225	0.7247 ± 0.3027
5gram	0.4430 ± 0.5087	0.4977 ± 0.3484

Table 4.2. General performance of the net trained on the cross-dependency grammar.

98% (decreasing it from 9.9283 to 0.1931) and the mean cosine performance improved by 125% (from $cos = 0.3597$ for the untrained net to $cos = 0.81$ after training). And, once more, it is clear that predictions based on relative word frequency are highly inadequate for the task at hand. Finally, the detailed behavior on complex structures evinced by the net trained using the incremental memory strategy again surpassed that of the adult net, even though the two nets displayed a quite similar general performance (adult net: MSE: 01978 $\pm$ 0.3461; mean cosine: 07950 $\pm$ 0.2482).

4.2.2 Performance on Non-iterative Recursive Structures

We saw above that both nets were able to achieve a level of performance above that of n-gram based predictions. A more detailed analysis of the latter revealed certain shortcomings compared with the nets. Predictions using n-grams are not able to deal with more complex structures where agreement features have to be remembered over long distances. The adult nets, on the other hand, were inconsistent in their predictions on complex structures. As was to be expected following Elman (1993), the maturationally constrained nets were much more consistent in their prediction behavior *vis-a-vis* complex grammatical regularities. Next, we shall see in detail how the nets dealt with multiple instances of center-embedding and cross-dependency, respectively.

Multiple Center-embeddings

In chapter 3, we saw that nets were able to learn a limited amount of center-embedding (mirror recursion), and furthermore exhibited a behavior similar to that of humans on similar structures. Recall that the mirror recursion simulations did not involve instances of i-recursion, so the nets could concentrate entirely on learning the former. However, in the human case, instances of ni- and i-recursion are interspersed with each other in the input that children receive. It therefore remains to be seen whether a net can learn

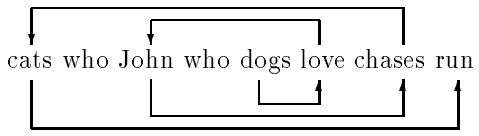


Figure 4.4. An illustration of the subject noun/verb agreement dependencies (arrows below) and the object nouns of the transitive verbs (the arrows above) in the doubly center-embedded sentence *'cats who John who dogs love chases run'*.

a limited degree of ni-recursion from input also incorporating i-recursive structures. Elman (1991b, 1993) has demonstrated that some ni-recursion can be achieved when a net is trained on input with center-embeddings and a single kind of i-recursion (in the form of right-branching subject relative clauses). In the present simulation based on the grammar in Figure 4.1., the net was faced with the much more complex task of learning ni-recursive regularities from input also containing several instances of right-branching i-recursion (sentential complements, prepositional modifications of NPs, conjunction of NPs, and subject relative clauses) as well as instances of left-branching i-recursion (prenominal genitives).

Figure 4.4 provides an illustration of the dependency structure involved in the processing of the multiple center-embedded sentence *'cats who John who dogs love chases run'*. Notice that the two object relative clauses require transitive verbs which have the two first nouns as their object (hence creating the double center-embedding). Figure 4.5 shows the prediction made by the net (in terms of summed activations) when processing each word in the above sentence[7].

In (a), we see the initial state of the network at the beginning of a sentence. Here the net is expecting either a singular or a plural noun. Having received *'cats ...'* in (b) the net predicts that it will get a plural verb, *'who'*, *'and'*, a preposition, or a singular genitive marker (the last three predictions are in `misc`). Next, the net anticipates receiving either a noun or a plural verb in (c). Given the context *'cats who John ...'* as in (d) the net correctly predicts a transitive singular verb, because it has realized that an object relative clause has begun (see Figure 4.4), another *'who'*, an *'and'*, a singular genitive marker, or a preposition starting a PP modification of John. The picture in (e) is similar to that of (c)—except that the net is predicting a singular verb. In (f), the net rightly presupposes that yet another object relative clause may have begun and

[7]In Figure 4.5 and 4.7, `s-N` refers to proper/singular nouns; `p-N` to plural nouns; `s-iV`, `s-tV`, and `s-cV` to singular intransitive, transitive verbs plus optionally transitive, and clausal verbs, respectively; `p-iV`, `p-tV`, and `p-cV` to plural intransitive, transitive verbs plus optionally transitive, and clausal verbs, respectively; `wh` to *who*; `eos` to end of sentence marker, and `misc` to *that*, *and*, genitive markers, prepositions, and the nouns used with the prepositions.

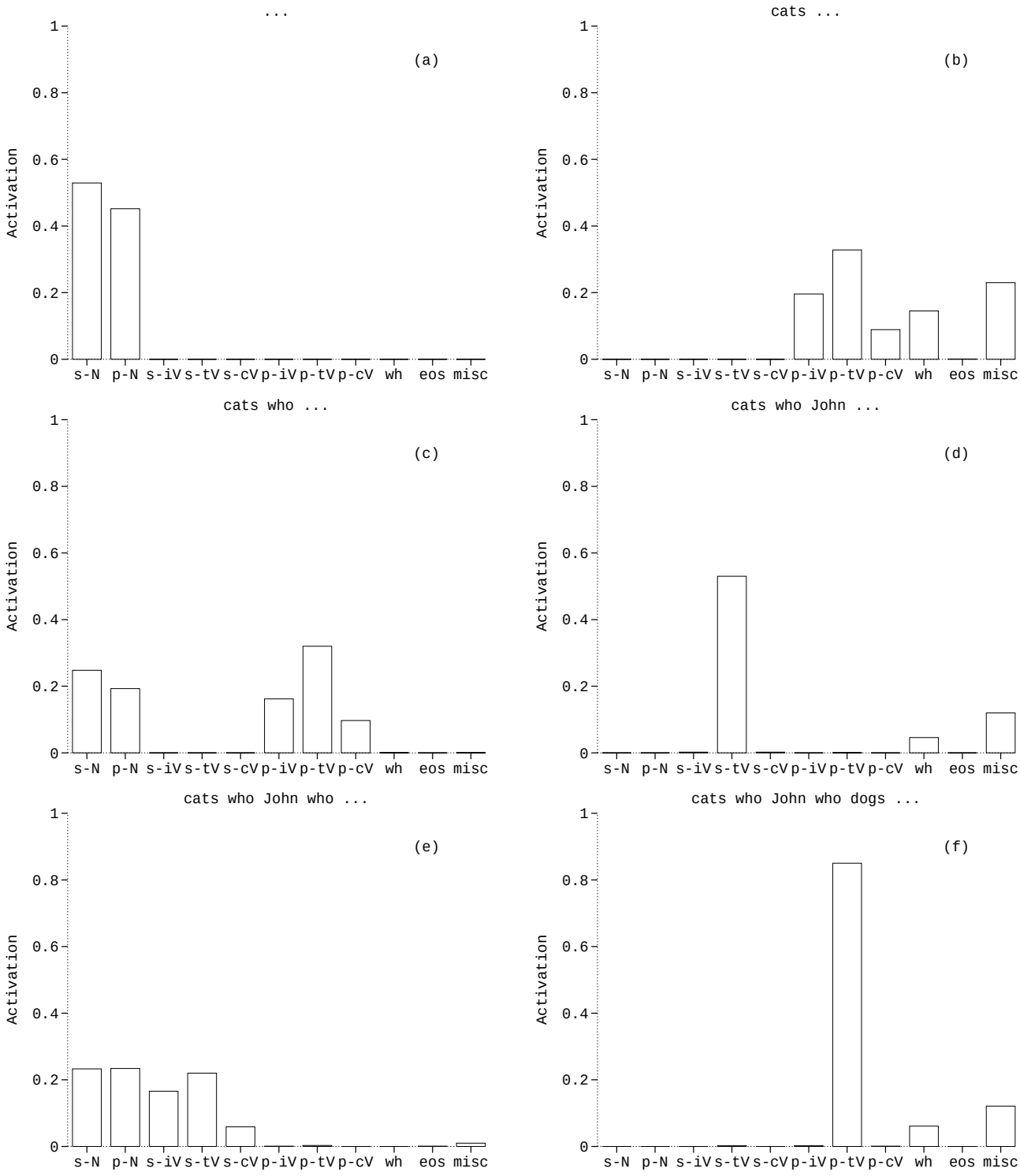


Figure 4.5. Network predictions after each word in the center-embedded sentence *'cats who John who dogs love chases run'*.

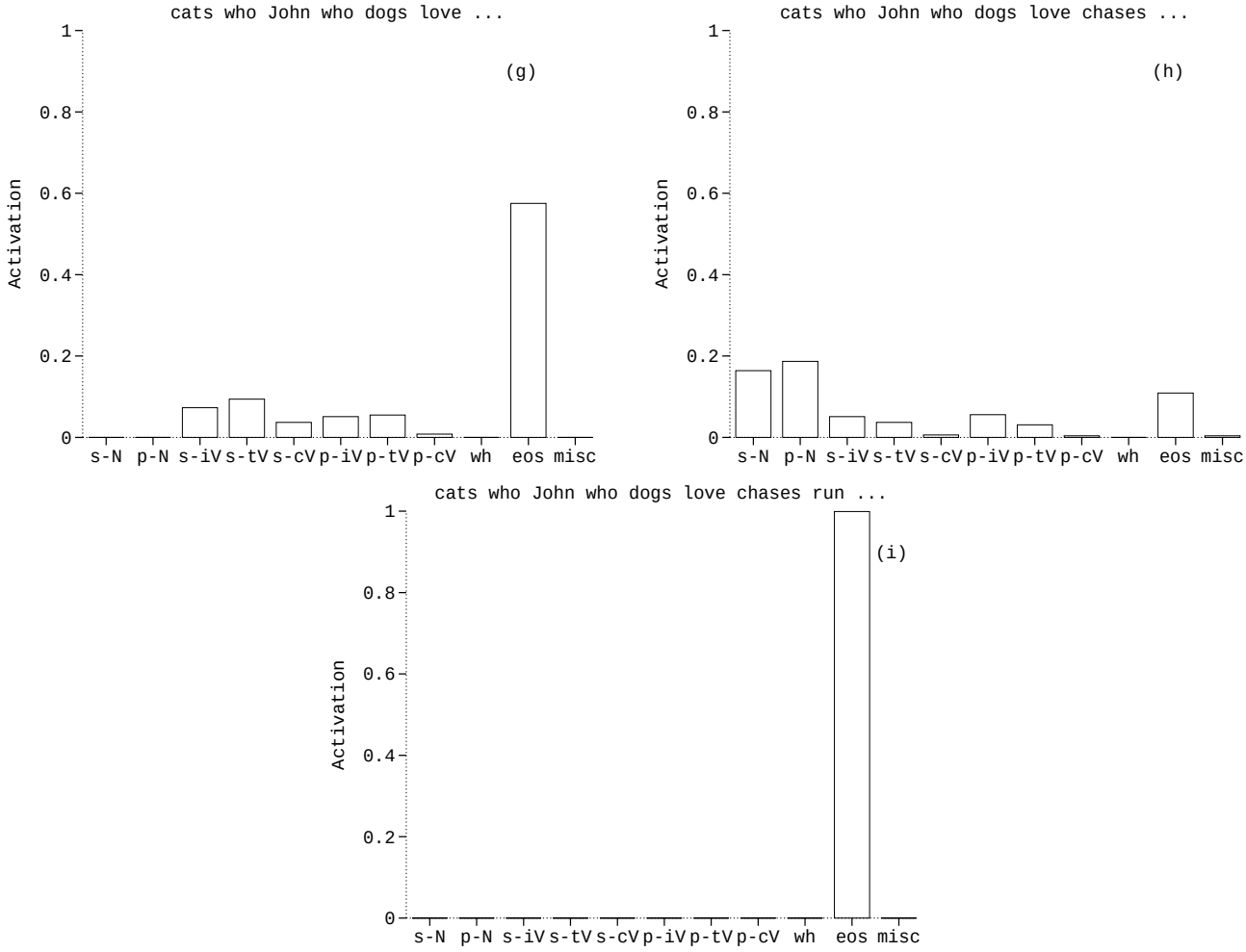


Figure 4.5. continued.

predicts a plural transitive verb to match with *'dogs*, or, alternatively, a third *'who'*, an *'and'*, a plural genitive marker, or a preposition. Problems arise in (g) when the net is supposed to predict the second verb whose subject noun is *'John'*. The net wrongly activates the end of sentence marker, the plural verbs as well as the intransitive and clausal forms of the singular verbs. The net should only have activated the transitive (and optionally transitive) singular verbs. That the latter have the highest activation of all the verb forms at least suggests that the net is going in the right direction. Again, in (h) we see that the net is somewhat off target. It should only have predicted plural verb forms, but also activated both nouns, singular verbs, and the end of sentence marker. Nevertheless, the net is able to correctly predict the end of sentence marker, once it has received the last verb, *'run'*, in (i).

The sudden breakdown of performance in (g) is not as detrimental for the net as a model of language acquisition and processing as one might initially think. In fact, this breakdown pattern closely follows the observed problems that humans have with similar sentences assessed in terms of recall (Miller & Isard, 1964), comprehension (even after some training on center-embedded structures, Larkin & Burns, 1977), and grammaticality judgements (Marks, 1968). Moreover, whereas the net had significant problems with doubly center-embedded sentences—as we saw above—it had no, or very little, trouble with sentences involving a single center-embedding. This has also been demonstrated in the human case (Bach, Brown & Marslen-Wilson, 1986; Larkin & Burns, 1977; Marks, 1968; Miller & Isard, 1964). This means that not only was the net able to reproduce the behavior observed on center-embedded structures in chapter 3 given considerably more complex input, but its performance also closely mimicked human processing behavior on the same sentence structures[8]. So, at least, when it comes to center-embedding, simple recurrent networks are viable candidates as models of human sentences processing. Next, we shall take a closer look at the detailed behavior of the network trained on the cross-dependency grammar from Figure 4.2.

Multiple Cross-dependencies

In contrast to the simulation involving the center-embedding grammar from Figure 4.1, there seems to be no precursors for the simulation incorporating the cross-dependency grammar, save the identity recursion simulations reported in the previous chapter.

[8]It should be noted, however, that this is not meant to suggest that humans would exhibit the exact same breakdown pattern, only that they will experience similar processing difficulties. For example, concerning Figure 4.5(g) this means that one cannot take the erroneous activation of `eos` as indicating that humans would predict end of sentence at this point. Instead, activations of illegal categories should be taken throughout this chapter as indications of processing difficulties.

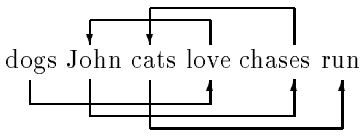


Figure 4.6. An illustration of the subject noun/verb agreement dependencies (arrows below) and the object nouns of the transitive verbs (the arrows above) in the sentence *'dogs John cats love chases run'* with two crossed dependencies.

The latter simulations demonstrated that recurrent networks were able to learn some degree of cross-dependency recursion, and, furthermore, display a behavior similar to that of humans on similar structure. This was not a trivial finding given that cross-dependencies are taken to require the power of context-sensitive languages. Nonetheless, as in the mirror recursion case, the training sets used to induce the regularities underlying identity recursion did not involve any instances of i-recursion. To allow a more close comparison with natural language, the present cross-dependency simulation additionally incorporates the same right- and left-branching i-recursive structures as in the center-embedding simulation.

Figure 4.6 shows the structure of the crossed subject noun/verb dependencies in the sentence *'dogs John cats love chases run'* as well as the object nouns of the two transitive verbs. Recall that this sentence structure is meant to correspond to Dutch with an English gloss. In Dutch, crossed dependencies allow the objects of transitive verbs to follow the latter's subject nouns. So, in Figure 4.6 we see *'John'* following the main subject noun *'dogs'* as the object of the main verb *'love'*. *'John'* is also the subject of the first subordinate clause with *'chases'* as its verb and with *'cats'* as its object. The latter is located as the third consecutive noun and is, in turn, the subject of a second subordinate clause having *'run'* as its verb. Figure 4.7 illustrates the behavior of the network when processing the above sentence involving two crossed dependencies.

The first histogram (a) shows that the net always expects a noun as the first word in a sentence. When it receives *'dogs'* in (b), it predicts that the next word is either another noun (leading to crossed dependencies), a plural verb of any form, *'who'*, *'and'*, a preposition, or a plural genitive marker (the last three predictions are collapsed in `misc`). Having seen *'dogs John ...'*, the net recognizes in (c) that it is receiving a sentence with, at least, one crossed dependency, and correctly anticipates either a noun or a plural transitive verb as the subsequent input. This pattern is replicated in (d) after the net is fed a third noun, *'cats'*. Given *'love'* as input in (e), the net rightly activates the singular transitive verbs to match with *'John'* (see Figure 4.6), but also erroneously activates the plural transitive verbs. Notice that the correct activations are twice as high as the incorrect activations. In (f) the erroneous activations have

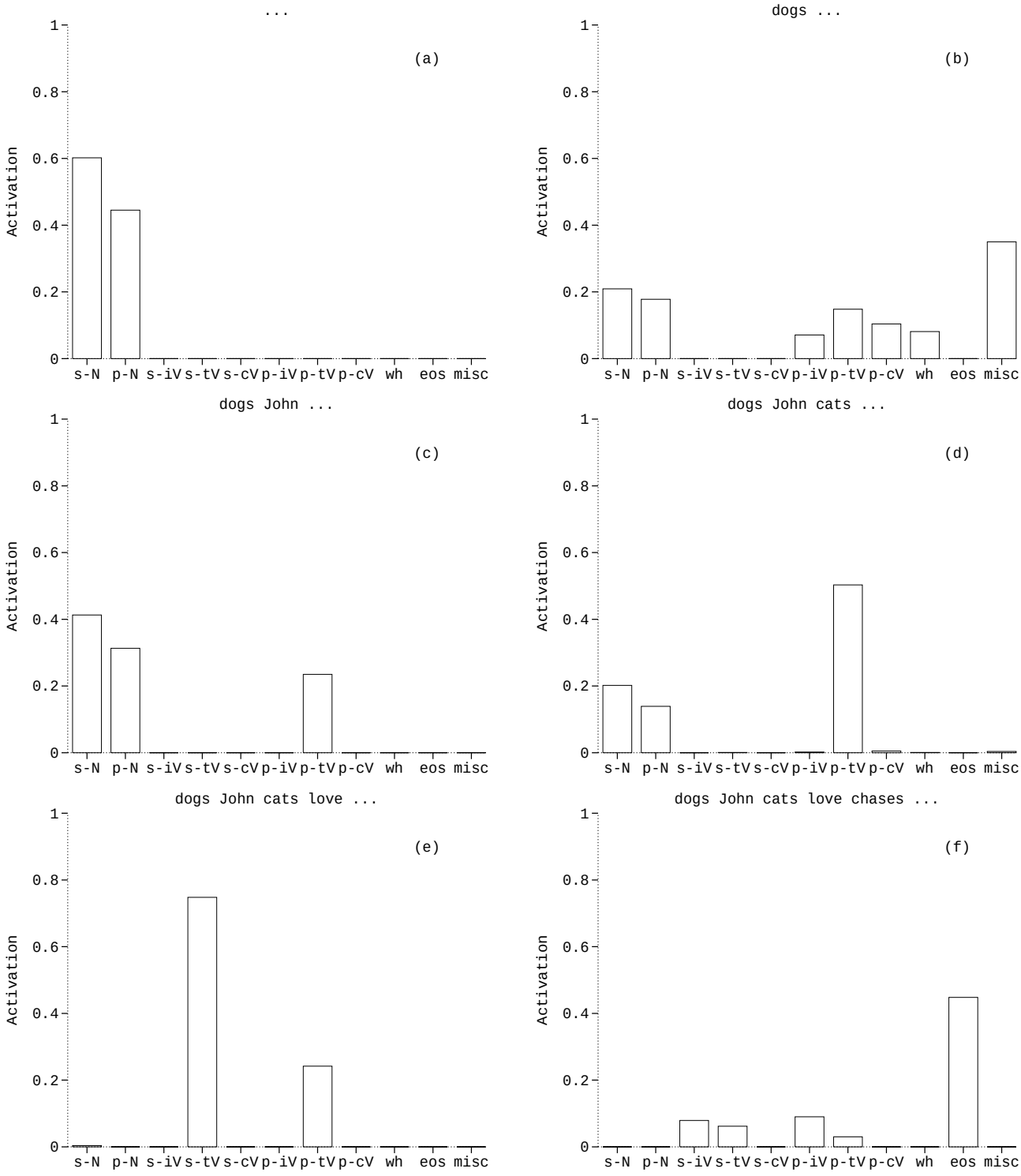


Figure 4.7. Network predictions after each word in the cross-dependency sentence *dogs John cats love chases run.'*.

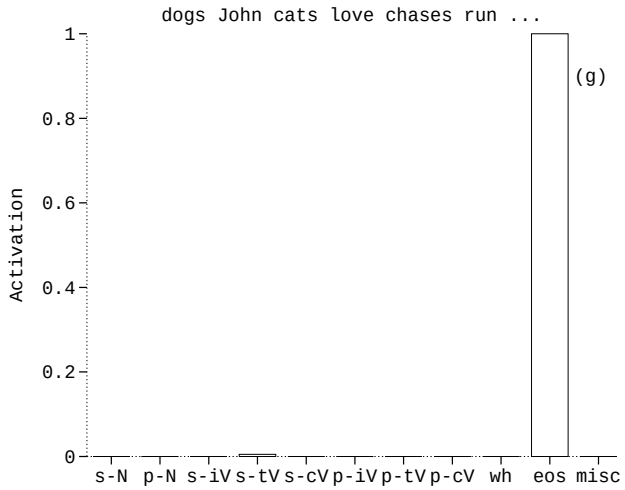


Figure 4.7. continued.

diminished. The net appropriately predicts a plural verb and the end of sentence marker, but wrongly activates singular verb forms. Following the last verb *'run'* in (g), the net confidently predicts end of sentence as it should.

As it was the case in the center-embedding simulation, the present network experiences problems when it is to predict the second verb in (e). However, the breakdown of performance here is not as severe as with the net trained on the center-embedding grammar. Moreover, the recovery from the problems following the second verb prediction also appears to be better. This difference could be a consequence of a heavier memory load in the processing of the center-embedded sentence because of the additional two instances of *'who'* before the first verb is encountered. No such difference was observed in the simulations presented in chapter 3. Still, it is worth noticing that humans appear to perform better on cross-dependency structures than on center-embedded constructions of similar depth (Bach, Brown & Marslen-Wilson, 1986). So, in this way the difference in performance between the net trained on the center-embedding grammar and the net trained on the cross-dependency grammar may reflect a real difference in the learnability and processability between the languages that the two grammars generate—at least, insofar as the same difference is also found in human performance. But, can the two nets perform equal to humans on i-recursive structures? I turn to this question next.

4.2.3 Performance on Iterative Recursive Structures

The results presented above suggested that the nets were able to deal with ni-recursion in a human-like manner. In chapter 2, I argued that a model of language not only should be able to process a limited degree of ni-recursion, but also must be able to handle a considerable amount of left- and right-branching i-recursion. The following four subsections report the nets' behavior on left-recursive structures exemplified by multiple prenominal genitives and right-recursive structures instantiated by multiple embeddings of sentential complements, multiple subject relative clauses, and multiple prepositional modifications of NPs. Since the two nets exhibited very similar behavior on these i-recursive structures, I only report two examples from each net as illustrations of their common level of performance.

Multiple Prenominal Genitives

Left-branching i-recursive structures are not frequent in English which is a predominately right-branching language (as opposed to, e.g., Japanese). The left-branching construction that presumably occurs most often in English is the prenominal genitive phrase. This construction permits the modification of nouns in a left-branching fashion as when we want to express the complex ownership of a given cat in terms of the phrase *'Bob's teacher's mother's cat'*. Figure 4.8 demonstrates the patterns of activation (in the cross-dependency trained net) when processing selected words in the sentence *'Mary's boys' cats' John sees'* with three prenominal genitives[9].

We have already seen the initial activation patterns of both nets in Figures 4.5 and 4.7. In (a), the net is predicting from the context of *'Mary ...'*, suggesting that the next input will be a noun, a singular verb, a singular genitive marker, a preposition, *'and'*, or *'who'* (the activation of the last three being collapsed in `misc`). Once the net receives the singular genitive marker, it expects that a noun must be next in (b). Given the plural noun *'boys'* in (c) the net predicts that the next word is either a plural verb or a plural genitive marker. When the net subsequently gets another genitive marker as its input, it activates singular and plural nouns only (not shown here, but similarly to (b)). Next, following the input of *'cats'* in (d), we have an activation pattern similar to (c)—albeit the activation is lower for both the plural verbs and the plural genitive marker, and a small erroneous activation of the end of sentence marker appears. The

[9]In the remaining figures in this chapter, the verb forms have been collapsed into a set of singular verbs `s-V` and a set of plural verbs `p-V`. The single and plural genitive marker have been separated from the `misc` group as, respectively, `s-g` and `p-g`. The group `misc` in Figure 4.8 covers activations of *that, who, and*, prepositions, and the nouns used with the latter.

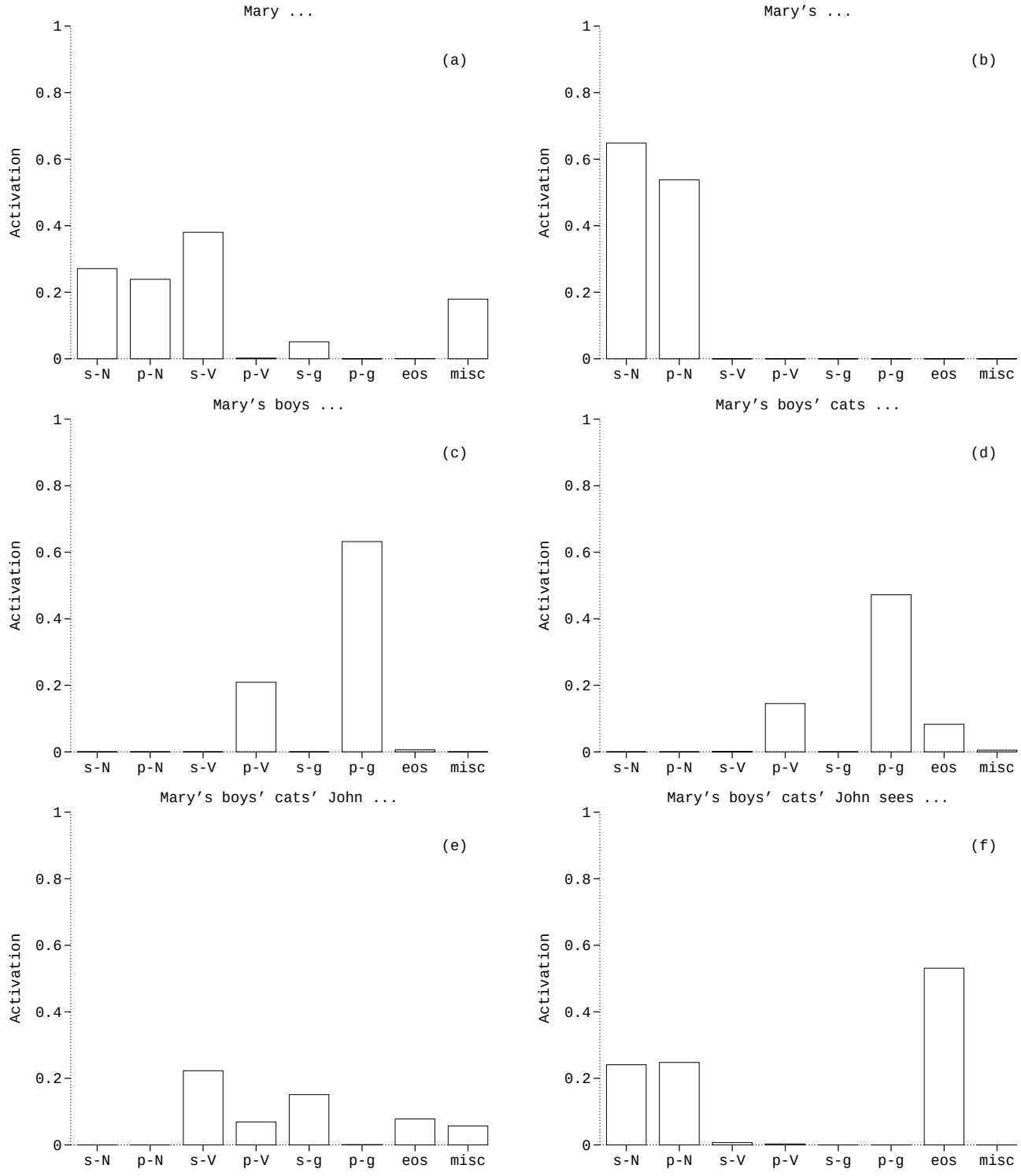


Figure 4.8. Network predictions after selected words in a sentence with multiple prenominal genitives: *'Mary's boys' cats' John sees'*.

net's problems grow somewhat in (e), showing the correct prediction of a singular verb, the singular genitive marker as well as a small activation of *'and'*, *'who'* and the prepositions, but also the incorrect prediction of a plural verb and, again, the end of sentence marker. The net gets back on track in (f), expecting either an object noun or an end of sentence marker after *'sees'*. Thus, it seems that the net is able to deal with left-branching i-recursion but in a, perhaps, non-optimal fashion given the apparent increase in error following the increase in recursion depth. I will return to this point later, and continue with an illustration of net behavior during the processing of multiple sentential complements.

Multiple Embeddings of Sentential Complements

Sentential complements provide a convenient way of expressing, for instance, propositional attitudes in English, such as, *'Mary thinks that cats chase dogs'*. Figure 4.9 depicts the activation patterns (in the net trained on the center-embedding grammar) observed when processing the sentence *'Mary says that men think that John knows that cats run'* which incorporates three sentential complements[10].

Having received *'Mary ...'* in (a), the net predicts that the next word will be either a singular verb, a singular genitive marker, a preposition, *'and'*, or *'who'* (with the last three predictions grouped together in `misc`). In (b), the net has received the clausal verb *'says'* and correctly predicts that the complementizer *'that'* must come next. After the complementizer has been given to the net as input, only a noun can follow, which is what the net predicts in (c). When the net receives the predicted noun, in this case *'men'*, the activation pattern displayed in (d) is similar to that of (a)—save the correct prediction of a plural verb matching the plural input noun. The subsequent five prediction patterns follow exactly the previous patterns in the order (b), (c), (a), (b), (c) (and they will therefore not be shown here). The pattern of summed activation as illustrated in (e) is similar to (d), but produced given the context of *'Mary says that men think that John knows that cats ...'*. Finally, we see in (f) that the net rightly predicts the end of sentence marker after the intransitive verb *'run'*. It is clear that the nets were better at processing sentential complements than prenominal genitives. In fact, there was no erroneous activation at all. This may suggest that sentential complements are also easier for humans to process compared with prenominal genitives. Leaving further discussion of this issue for later, I now describe the nets ability to process multiple object relative clauses.

[10]Note that the complementizer *'that'* has been separated from the `misc` group. Otherwise, the labels in Figure 4.9 are the same as in the previous figure.

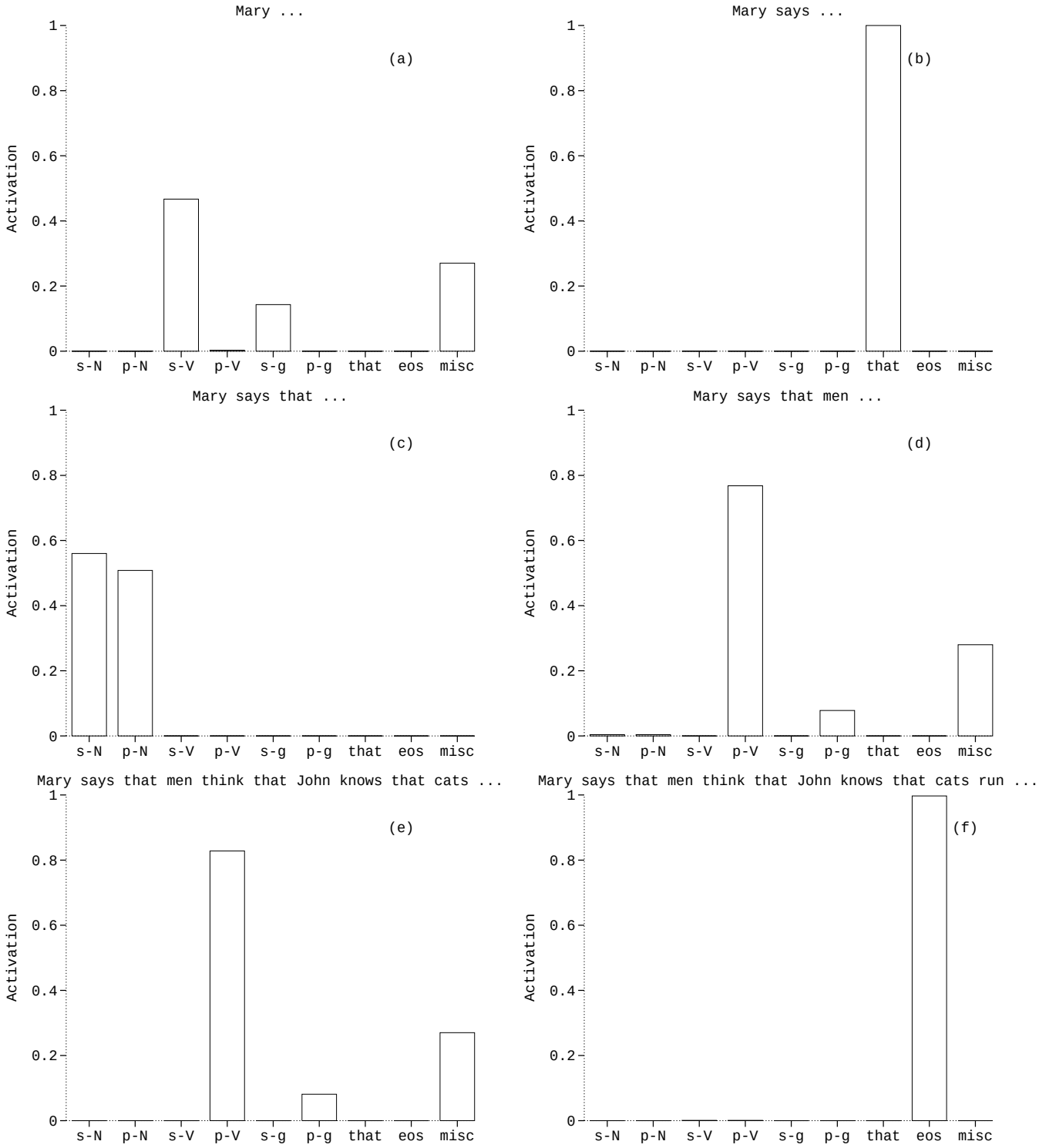


Figure 4.9. Network predictions after selected words in a sentence with multiple sentential complements: *'Mary says that men think that John knows that cats run'*.

Multiple Right-embedded Relative Clauses

As mentioned earlier, right-embedded relative clauses allow us to rephrase the content of both center-embedded and cross-dependency sentences in a more readily understandable form. Thus, both the center-embedded and the cross-dependency sentences above can be rewritten as *'dogs love John who chases cats who run'*. Numerous studies have demonstrated that multiple subject relative clauses are considerably easier to process than center-embedded sentences (Blaubergs & Braine, 1974; Foss & Cairns, 1970; King & Just, 1991; Marks, 1968; Miller & Isard, 1964) and cross-dependency sentences (Bach, Brown & Marslen-Wilson, 1986) of a similar depths of recursion. Figure 4.10 shows the behavior of the (cross-dependency grammar trained) network whilst processing the sentence *'dogs love John who chases cats who see Mary who chases girls'* involving three right-embedded relative clauses[11].

First, the pattern of summed activation pattern given the context *'dogs love ...'* is presented in (a), depicting an activation of the nouns as objects of the transitive verb. Next, the net predicts either *'who*, the singular genitive marker, end of sentence, a preposition, or *'and'* (the last two collapsed in the `misc` group). The net also has an incorrect, but insignificant, activation of the verbs. In (c), we also find a very small erroneous activation of the plural verbs which is dwarfed by the correct activation of the singular verbs. When the net subsequently gets *'chases'* as input, it again predicts only nouns as in (a). Having received *'cats'* in (d), the net produce a pattern of activation similar to (b), but with the plural genitive marker activated instead of the singular one. Given a second *'who'* in (e), the net rightly predicts a plural verb with some minor incorrect activations of the singular verbs. Presented with the optionally transitive verb *'see'* in (f), either an object noun or an end of sentence marker is correctly predicted by the net. The previous erroneous activation have now become somewhat higher as exhibited by the activations of the plural and the singular verb forms. As we can see from (g), the net displays some spurious, but minute, activations of both nouns, verbs, and the plural genitive marker, all of which are not allowed in the context of *'dogs love John who chases cats who see Mary ...'*. Still, it by far rightly predicts *'who'*, a singular genitive marker, end of sentence, a preposition, and a conjunction. The prediction pattern in (h) is similar to that of (c), although the singular verb group is activated less strongly and the wrong activations slightly more pronounced this time. A comparison between the noun predictions in (a) and in (i)—both following transitive verbs—indicates that the net has become less confident in its predictions as the number

[11] In this figure, *'who'* has been separated from the *misc* group. The remaining labels correspond to those of Figure 4.8.

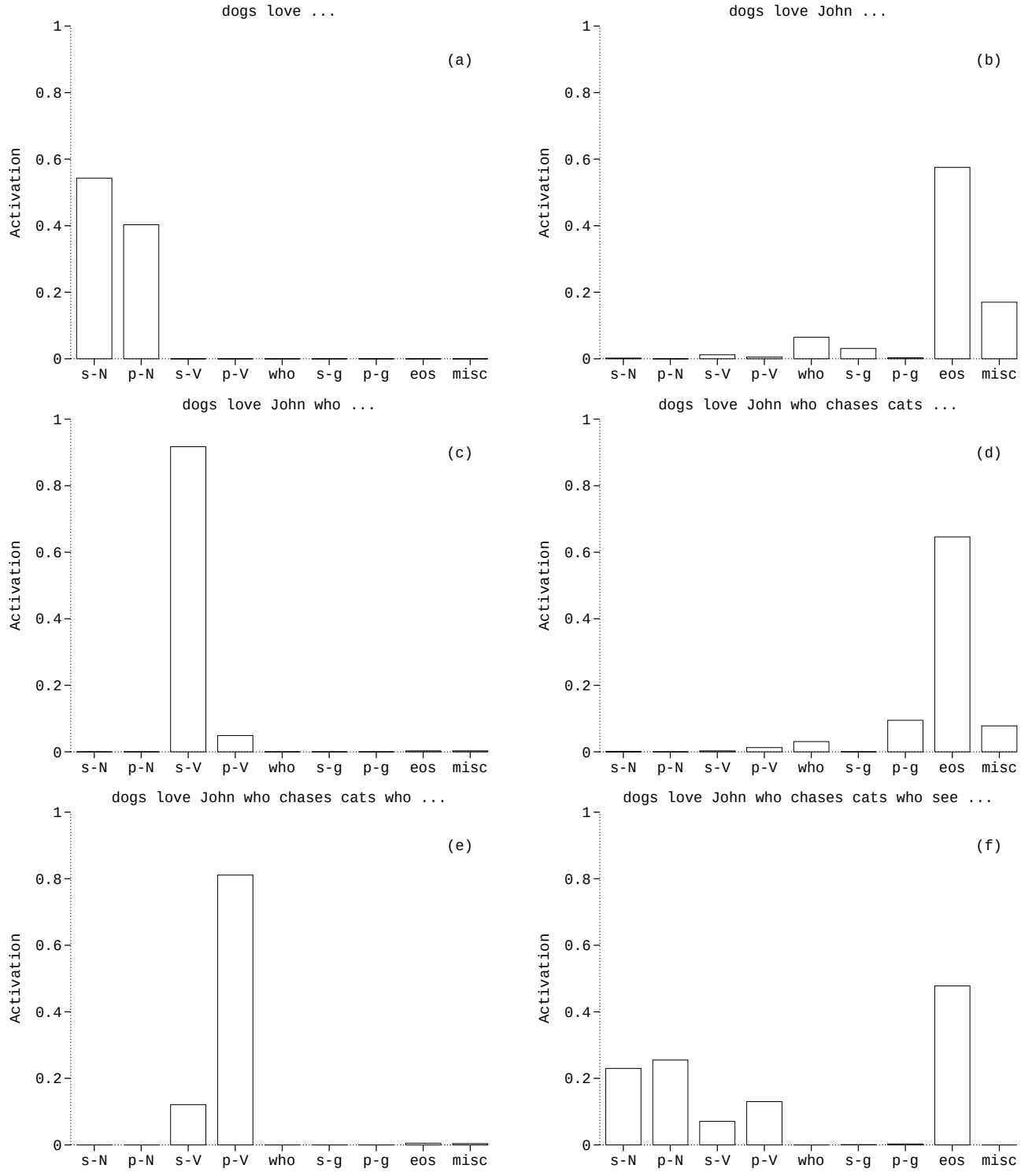


Figure 4.10. Network predictions after selected words in a sentence with multiple right relative clauses: *'dogs love John who chases cats who see Mary who chases girls'.*

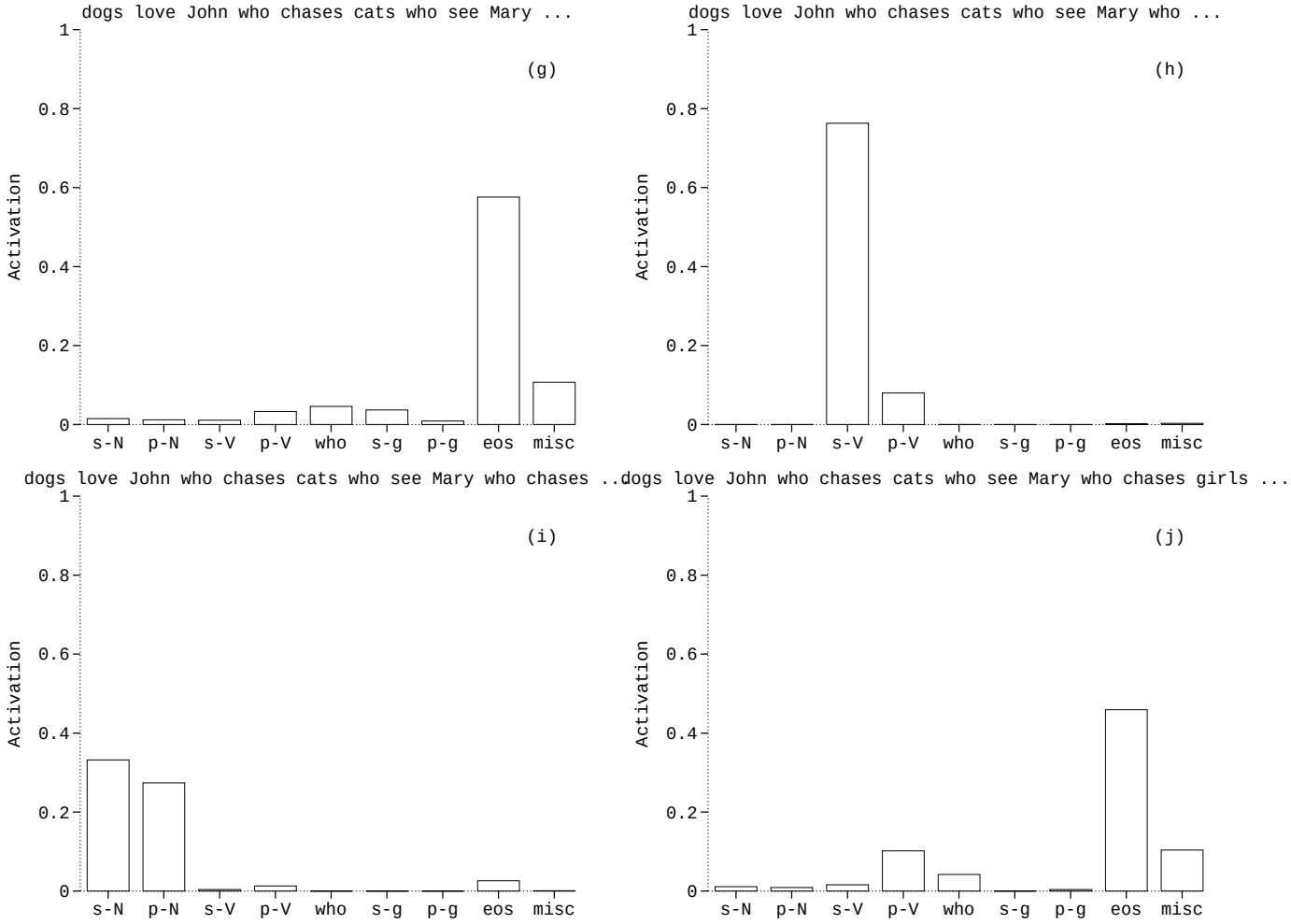


Figure 4.10. continued.

of embeddings increase. The predictions shown in the last histogram (j) continues this tendency with spurious, mistaken activation of the nouns and verbs, and decreased confidence in the correct predictions of *'who'*, end of sentence, and the prepositions. It thus seems that net performance on subject relative clauses, though superior to performance on ni-recursive structures of a similar degree of embedding, nonetheless show evidence of degradation as sentence length increases. I will return to this after outlining net behavior on sentences involving multiple prepositional modifications of an NP.

Multiple Prepositional Modifications of NPs

The final example of i-recursion in the present simulations addresses the processing of complex NPs, such as, *'the flowers in the vase on the table near the box'*. Such multiple instances of prepositional modifications of NPs are notorious for their inherent ambiguity. For instance, is it the vase or the table which is near the box? Leaving these ambiguity issues aside here (they also concern ambiguity concerning NP or VP attachment of the PP), I focus on simple right-branching PPs modifying a noun. Figure 4.11 demonstrates the behavior of the (center-embedding trained) network during the processing of the sentence *'cats from lake in town near city chase girls'* incorporating two PP embeddings[12].

Given the context of *'cats from ...'* in (a), the net confidently predicts that the following word must be one of the nouns used with the prepositions. Next, in (b) the net rightly expects either a plural noun or another preposition, but also produce a small amount of erroneous activation of the singular verbs and the end of sentence marker. As in (a), the net has no problems predicting another preposition noun subsequent to receiving the preposition *'in'* (not shown). In (c), the pattern of (b) repeats itself with the correct activations of plural verbs and prepositions, but with some mistaken, minor activation of the singular verbs and the end of sentence marker. This pattern of error becomes slightly more evident in (d), which also displays a reduced confidence in the right predictions given the context *'cats from lake in town near city ...'*. Having received *'chase'*, the verb that the net was waiting for, it recovers and predicts that a noun must come next as the obligatory object of transitive input verb. Finally, the net makes the appropriate predictions in (f) following *'dogs'*; that is, either the next word is a preposition, a plural genitive marker, end of sentence, *'who'*, or *'and'*. As with the right-branching relative clauses exemplified in Figure 4.10, there is a tendency for the

[12]In Figure 4.11, the prepositions and their nouns have been separated from the *misc* group. The remaining labels correspond to those of Figure 4.8.

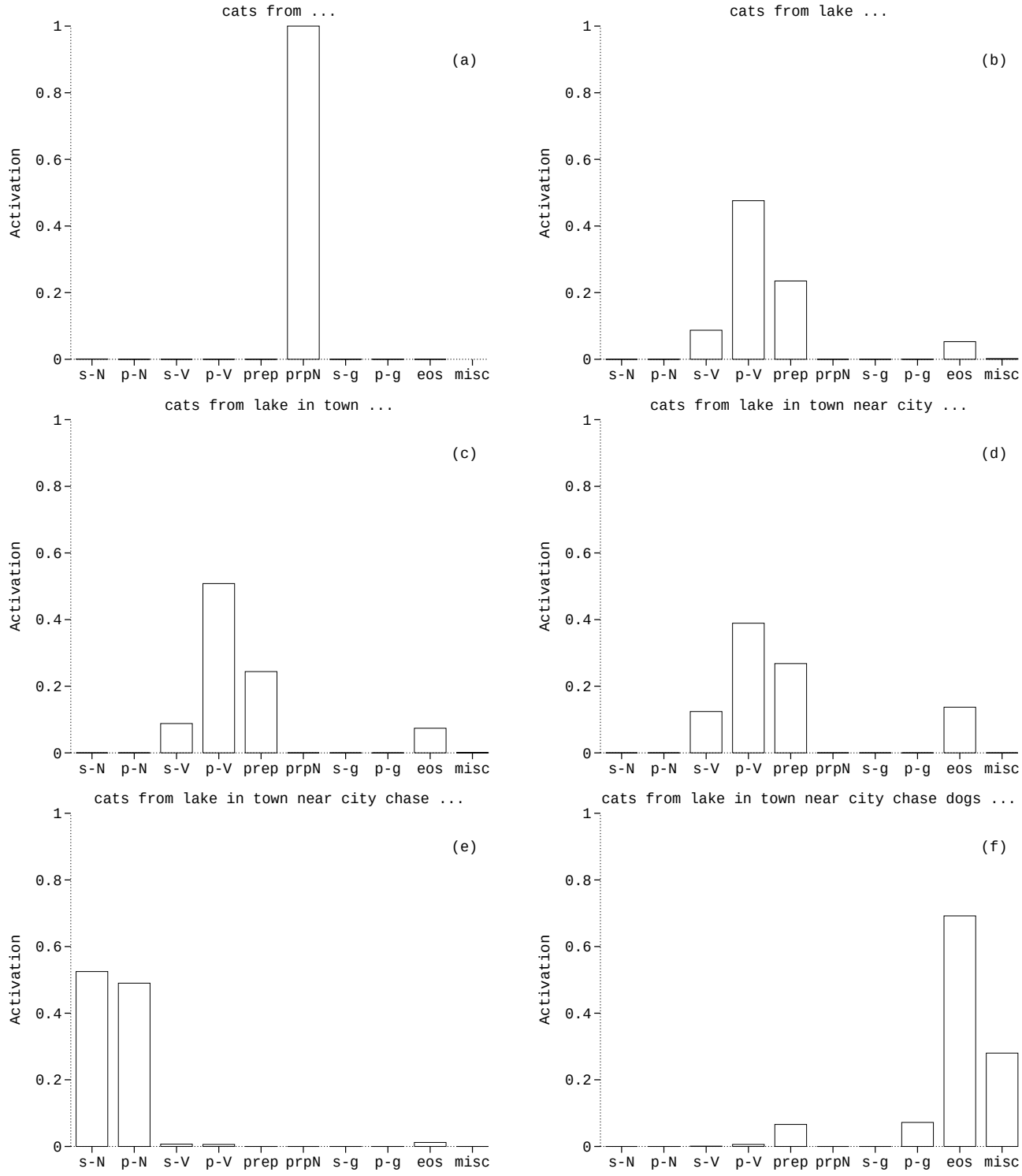


Figure 4.11. Network predictions after selected words in a sentence with multiple PP modifications of the subject noun, *'cats from lake in town near city chase girls'*.

performance to degrade as the degree of i-recursive complexity increases.

Given the patterns of slow degradation of performance on all the i-recursive structures, save the case of multiple sentential complements, the question is whether these patterns are comparable with observed human limitations on similar structures. We have already seen that net performance on both center-embedded and cross-dependency structures closely mimicked human behavior on the same ni-recursive constructs. It is typically assumed that humans are able to deal with i-recursive structures of an in principle unbounded length. In classical parsers, incorporating grammars consisting of a recursive set of rules and some kind of stack memory structure (as described in chapter 2), both left- and right branching i-recursive structures of an infinite length can be processed—even when a limit is imposed on the stack depth. This is because i-recursion allows the parser to clear its memory at the start of each recursive level. The stack memory is therefore not in danger of being exhausted as, for example, in the case of center-embedded structures (see Figure 2.3 in chapter 2, for more details on the latter).

It is, however, not clear that humans have such an infinite capacity for i-recursion. Indeed, it appears (from a literature search and communication with colleagues) that human performance on i-recursive structures have not been studied explicitly. That is, the purported human ability to deal with an unlimited length of i-recursion has, to my knowledge, never been demonstrated in an experimental context. In fact, evidence exists pointing in the opposite direction and, in part, corroborating the simulation results on subject relative clauses as illustrated in Figure 4.10. When comparing sentences containing multiple center-embeddings with sentences consisting of multiple subject relative clauses, Blaubergs & Braine (1974) found that the comprehension of the latter also decreased (almost) proportionally with the increase in i-recursive complexity (contrasted with the more dramatic decrease in the performance on center-embedded sentences). This is indirectly comparable with the simulation results presented on the processing of multiple subject relative clauses. If this connection between network performance and human performance on i-recursive structures is genuine, then we may expect human performance on multiple instances of both prenominal genitives and prepositional modifications of NPs to follow the same kind of degradation relative to sentence complexity. Intuitively, this seems to be the case, but it remains to be tested experimentally.

This still leaves the perfect network performance on the multiple sentential complements, as depicted in Figure 4.9, to be explained. Two explanations, at least, comes to mind. It could be the case that this kind of i-recursive structure is easier for the

nets to process because the two grammars allow fewer prediction choices during the processing of these sentences. For example, only *'that'* can follow a clausal verb, and only nouns are permitted after this complementizer. In contrast, at most points in the processing of the other three i-recursive structures, more than one category is allowed. Human performance on sentential complements might thus be worse than what the nets exhibited because in English *'that'* can indicate either the start of a relative clause (as in *'I saw the house that you grew up in'*) or the beginning of a sentential complement (as it is used in the sentence in Figure 4.9). Alternatively, the simulation results could be taken to suggest that we may find a qualitative difference in human performance on sentences with multiple sentential complements compared with sentences involving numerous instances of the other three kinds of i-recursion discussed here. The bottom-line is that the simulation results from the processing of i-recursive structures make certain predictions which can be tested experimentally[13].

4.3 Generalization in Connectionist Networks

The simulation results reported above have suggested that simple recurrent networks are viable models of language processing. It remains to be seen whether they afford the kind of generalization abilities that we would expect from models of language. Recently, Hadley (1994a) has attacked connectionist models of language learning for not achieving a sufficient degree of generalization. He rightly points out that generalization in much connectionist research has not been viewed in a sophisticated fashion. Testing is typically performed by recording network output given a test set consisting of items not occurring in the original training set (as it, admittedly, was the case in experiment 2 in chapter 3). Hadley, in effect, criticizes connectionists for not going beyond using training and test sets which have been put together according to convenience. He therefore challenges connectionists to adopt a more methodological training and testing regime. I have already addressed this challenge elsewhere (in Christiansen & Chater, 1994—but see also the reply by Niklasson & van Gelder, 1994, and Hadley, 1994b, for a response to both replies), and I report and extend those results here[14].

[13]Importantly, the kind of performance degradation reported here was found to be a general trend throughout my exploratory simulations using nets of varying sizes and different configurations (such as, the auto-associative simple recurrent network of Maskara & Noetzel, 1992, 1993).

[14]The discussion of of Hadley's (1994a) notion of syntactic position and the formalization thereof is based on Christiansen & Chater (1994). The simulation results presented here are new and differ somewhat from the ones reported in the latter paper.

4.3.1 Degrees of Systematicity

In an effort to operationalize Fodor & Pylyshyn's (1988) abstract criticism of connectionism, Hadley (1994a) has defined different degrees to which a language learning system can generalize from experience, what he calls different degrees of systematicity. In his paper, he focuses on *syntactic* generalization, presenting notions of weak, quasi- and strong systematicity as benchmarks for connectionist models ('c-nets' in Hadley's terminology). As we shall see below, these definitions are rather vague insofar as more complex grammatical structure is concerned—as, e.g., in the simulations reported above. Despite Christiansen & Chater's (1994) criticism of this fact, and Hadley's (personal communication) acknowledgement of it, these definitions are reiterated in Hadley (1994b). Before we turn to Christiansen & Chater's discussion, and subsequent formalization, of Hadley's notions of syntactic systematicity, a brief look at his definitions are in order.

According to Hadley (1994a) "a c-net exhibits at least *weak systematicity* if it is capable of successfully processing (by recognizing or interpreting) novel test sentences, once the c-net has been trained on a corpus of sentences which are *representative*" (p. 6). A training corpus is 'representative' if "*every* word (noun, verb, etc.) that occurs in some sentence of the corpus also occurs (at some point) in every permissible syntactic position" (p. 6). *Quasi-systematicity* can be ascribed to a system if "(a) the system *can* exhibit at least weak systematicity, (b) the system successfully processes novel sentences containing embedded sentences, such that both the larger containing sentence and the embedded sentence are (respectively) structurally isomorphic to various sentences in the training corpus, (c) for each successfully processed novel sentence containing a word in an embedded sentence (e.g., 'Bob knows that Mary saw *Tom*') there exists some *simple* sentence in the training corpus which contains that same word in the same syntactic position as it occurs within the embedded sentence (e.g., 'Jane saw *Tom*')" (p. 6–7). Finally, a system will exhibit *strong systematicity* if "(i) it *can* exhibit weak systematicity, (ii) it can correctly process a variety of novel *simple* sentences and novel *embedded* sentences containing previously learned words in positions where they *do not appear* in the training corpus (i.e. the word within the novel sentence does *not appear in that same syntactic position* within any *simple or embedded* sentence in the training corpus)" (p. 7).

Central to each definition is the notion of *'syntactic position'*, which may or may not be shared between items in the training and test sets. Since syntactic position is not a standard term in linguistics, and since it is not discussed in either of Hadley's (1994a, 1994) papers, it is necessary to examine his examples to discover what meaning

is intended. These are concerned with the relationship between verbs and their arguments. The various argument positions of a verb (subject, direct object and indirect object) are taken to count as distinct syntactic positions. Also, the active and passive forms of a verb are taken to occupy different syntactic positions.

If these examples are taken at face value, difficulties emerge. For example, a lexical item is the subject with respect to some verb whether or not it occurs within an embedded sentence, a simple sentence, or the main clause of a sentence which contains an embedded sentence (and similarly with the other examples). This means that, for Hadley, *'John'* has the same syntactic position in *'John loves Mary'* as in *'Bill thinks that John loves Mary'*—indeed, this is explicit in point (c) of the definition of quasi-systematicity. Nonetheless, it would appear that, according to Hadley, a learning system which generalizes from either of these sentence to the other only requires weak systematicity (since no item occurs in a novel syntactic position). Yet, this seems to be exactly the kind of case which is supposed to distinguish quasi-systematicity from weak systematicity in Hadley's definitions. But, as we see, it appears that weak systematicity already deals with such cases, if syntactic position is defined in terms of grammatical role, since grammatical role abstracts away from embedding. Quasi- and weak systematicity therefore appear to be equivalent.

Presumably, either weak or quasi-systematicity is intended to have an additional condition, which is not explicit in Hadley's definition. The suggestion is made below that quasi-systematicity is only exhibited when the test and training sets contain embedded sentences. An alternative interpretation would be that Hadley is implicitly making use of a more global notion of syntactic context, which distinguishes the syntactic position of a subject in a sentence which contains an embedded clause, and one that does not, for example[15].

In order to extend the account beyond the cases of subject and object, a more general account of syntactic position is needed. Christiansen & Chater (1994) have suggested a possible definition which is presented below. This definition, in turn, allows them to define what they call three levels of *generalization*, which are intended to be close to the spirit of Hadley's original definitions of systematicity.

[15]Hadley (personal communication) seems to lean towards the latter interpretation in a recent revision of his definition of weak systematicity: "the training corpus used to establish weak systematicity must present every word in every syntactic position and must do so at all levels of embedding found in the training and test corpus. In contrast, a quasi-systematic system does not have to meet the condition in the second conjunct, but does satisfy the first conjunct". Notice that this revision suggests that Elman's (1989, 1991a) net might be quasi-systematic after all (pace Hadley, 1994a, p. 17). Interestingly, in Hadley (1994b) the definition of quasi-systematicity is left out.

4.3.2 Syntactic Context

The syntactic position of a word is defined in terms of the phrase structure tree assigned to the sentence in which it occurs. Christiansen & Chater use phrase structure trees since they are linguistically standard and can be used in a precise and general way. No theoretical commitment to phrase structure based approaches to linguistic theory are intended there, nor is it here. This account could be given equally well in alternative linguistic frameworks.

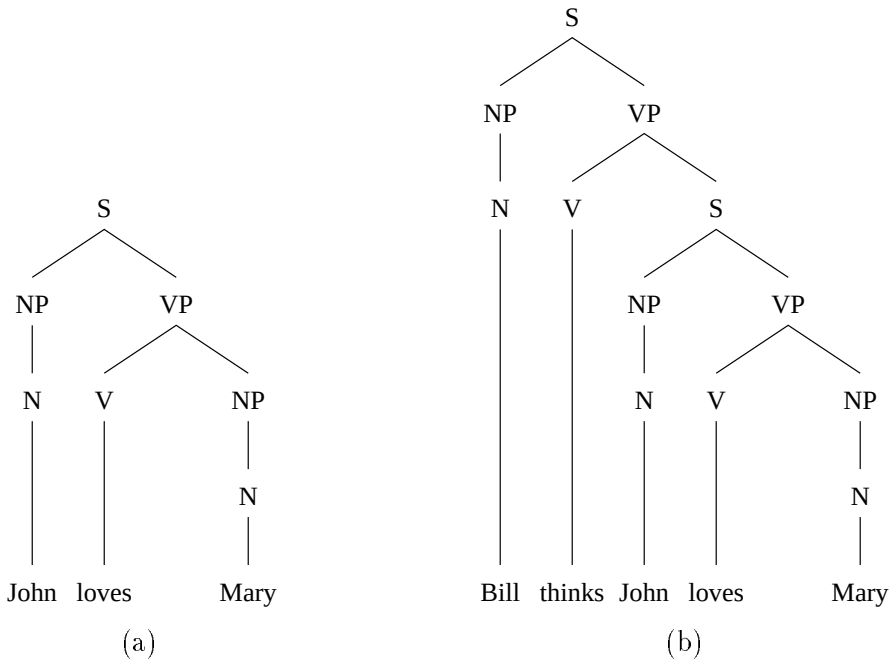


Figure 4.12. Phrase structure trees for (a) the simple sentence *'John loves Mary'* and (b) the complex sentence *'Bill thinks John loves Mary'*.

The syntactic position of a word is defined to be the tree subtended by the immediately dominating S or VP node, annotated by the position of the target word within that tree. This tree will be bounded below either by terminal nodes (Det, Proper Noun, etc.), or another S or VP-node (i.e., the syntactic structure of embedded sentences or verb phrases is not expanded). For example, consider the phrase structure trees for the simple sentence *'John loves Mary'* and the complex sentence *'Bill thinks John loves Mary'* as shown in Figure 4.12. In a simple sentence like (a), the subject is defined by its relation to the dominating S-node. The object and the verb are defined in relation to the verb phrase. This captures the distinction between subject and object noun positions. Figure 4.13(a) and (b) depict this distinction, illustrating, respectively, the syntactic positions of *'John'* and *'Mary'*.

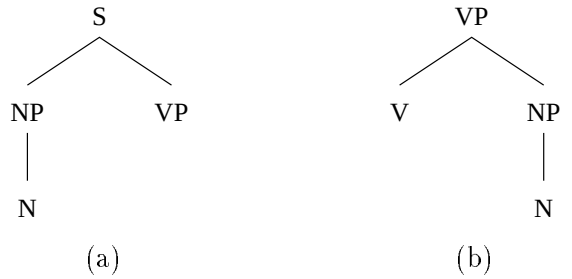


Figure 4.13. The syntactic position of (a) the subject noun and (b) the object noun in the sentence *'John loves Mary'*.

Also according to this definition, verbs with different argument structure are considered to have different syntactic contexts. For example, intransitive, transitive and ditransitive occurrences of verbs will be viewed as inhabiting different contexts. Furthermore, verb argument structure is relevant to the syntactic context of the object(s) of that verb, but not of its subject. In a complex sentence like 4.12(b), there will be different local trees for items in the main clause or in any embedded clauses. For example, *'thinks'*, which occurs in the main clause of 4.12(b), has a syntactic position defined with respect to the verb phrase pictured in Figure 4.14(a), whereas for *'loves'* in the embedded clause, the syntactic position is defined with respect to the structure of the embedded sentence shown in 4.14(b). The two trees in Figure 4.14 are thus examples of how the verb argument structure affects syntactic position. Notice that this means that the syntactic position within an embedded clause is affected only by its local context, and not by the rest of the sentence. Thus the notion of syntactic position applies independently of the depth of embedding at which a sentence is located. Furthermore, according to this definition, the syntactic context of a word in a particular clause in not affected by the structure of a subordinate clause; and the syntactic context of a word in an subordinate clause is not affected by the structure of the main clause.

Figure 4.14. The syntactic position of (a) the main verb and (b) the subordinate verb in the sentence *'Bill thinks John loves Mary'*.

4.3.3 Degrees of Generalization

Using this definition of syntactic position, Christiansen & Chater recast Hadley's definitions to give three levels of generalization for language learning systems[16].

i. *Weak Generalization:* A learning mechanism weakly generalizes if it can generalize to novel sentences in which *no* word occurs in a novel syntactic position (i.e., a syntactic position in which it does not occur during training)[17].

ii. *Quasi-Generalization:* A learning mechanism is capable of quasi-generalization if it can generalize to novel sentences as in (1), with the additional constraint that embedding must occur in the grammar.

iii. *Strong Generalization:* A learning mechanism strongly generalizes if it can generalize to novel sentences, that is, to sentences in which some (sufficiently many) words occur in novel syntactic positions. It is furthermore required that embedding occurs in the grammar, and that the learning mechanism can strongly generalize to both simple and embedded sentences[18].

Given this definition of strong generalization, consider the following two test sentences:

John thinks Bill loves Mary.
Bill loves Mary.

If *'Mary'* does not occur in the object position in the training set (in either embedded or main clauses), the syntactic position of *'Mary'* in both these sentences is novel. Thus, for a net to be ascribed strong generalization, it is necessary that it be able to process both sentences. On the other hand, if *'Mary'* did occur in object position even just once in the training set, then in neither of the two sentences is the syntactic position novel (and the net can therefore, at most, be characterized as capturing quasi-generalization). Thus, the above definitions are meant to capture the spirit of Hadley's (1994) proposals

[16]Note that further formalization may perhaps be needed to capture the full complexity of natural language. However, this would presumably have to take place within a given linguistic framework at the cost of the inter-theoretical compatibility sought for in Christiansen & Chater (1994).

[17]Note that Hadley's revised definition of weak systematicity (as mentioned in a previous footnote) differs from this notion of weak generalization.

[18]The requirement concerning simple and embedded sentences has been added to the definition found in Christiansen & Chater (1994) following worries expressed in Hadley (1994b). He criticized the definition of strong generalization for being easier to meet than his notion of strong systematicity, since the latter definition requires that the net be able to process both simple and embedded novel sentences. As it were, the example of strong generalization presented in Christiansen & Chater does meet the requirements of this new definition.

in a reasonably precise and general way. Next, I present some simulation results which aim to test how readily these definitions can be met by a simple recurrent network.

4.3.4 Generalization Results

As a first step towards meeting the strong generalization criterion described above, I report additional results from the simulation involving the cross-dependency grammar depicted in Figure 4.2. Results from the net trained on the center-embedding grammar illustrated in Figure 4.1 were presented in Christiansen & Chater (1994) and will not be treated in detail here. As mentioned earlier, both simulations build on and extend Elman's (1988, 1989, 1990, 1991a, 1991b, 1992, 1993) work on training simple recurrent networks to learn grammatical structure. However, Hadley (1994a) rightly notes that the training regime adopted by Elman does not afford strong systematicity (nor does it support the notion of strong generalization) since the net by the end of training will have seen all words in all possible syntactic positions. To address the issue of generalization, I therefore imposed an extra constraint on two of the nouns from Figure 4.3 (in both their singular and plural form). Thus, I ensured that *'girl'* and *'girls'* never occurred in a genitive context (e.g., neither *'girl's cats'* nor *'Mary's girls'* were allowed in the training set), and that *'boy'* and *'boys'* never occurred in the context of a noun phrase conjunction (e.g., both *'boys and men'* and *'John and boy'* were disallowed in the training corpus). Given these constraints I was able to test the net on known words in novel syntactic positions as required by the definition of strong generalization and by Hadley's notion of strong systematicity[19].

Strong Generalization in Genitive Context

Recall that neither *'girl'* nor *'girls'* has occurred in a genitive context in any of the training sets. Figure 4.15 illustrates the behavior of the net when processing the sentence *'Mary's girls run'* in which the known word *'girls'* occupies the novel syntactic position constituted by the genitive context (and the control sentence *'Mary's cats run'*)[20].

We have already seen evidence of net processing up to the point of *'Mary's ...'* in Figure 4.8(a) and (b)[21]. Here, in Figure 4.15(a), the behavior of the network is

[19]Hadley (personal communication) has acknowledged both test cases as possible *single* instances of strong systematicity; though these instances might not be sufficient to warrant the *general* ascription of strong systematicity to the net as a whole.

[20]The labels are the same as in Figure 4.8.

[21]It should be noted that the net was not able to predict neither *'girl'* nor *'girls'* after the genitive marker in *'Mary's ...'* (Figure 4.8(b)). At first, this would seem to preclude the ascription of strong generalization altogether, a point Hadley (1994b) has stressed. Christiansen & Chater (1994) treat this as a partial error, but I will argue below that it is an unimportant one and may, perhaps, even be

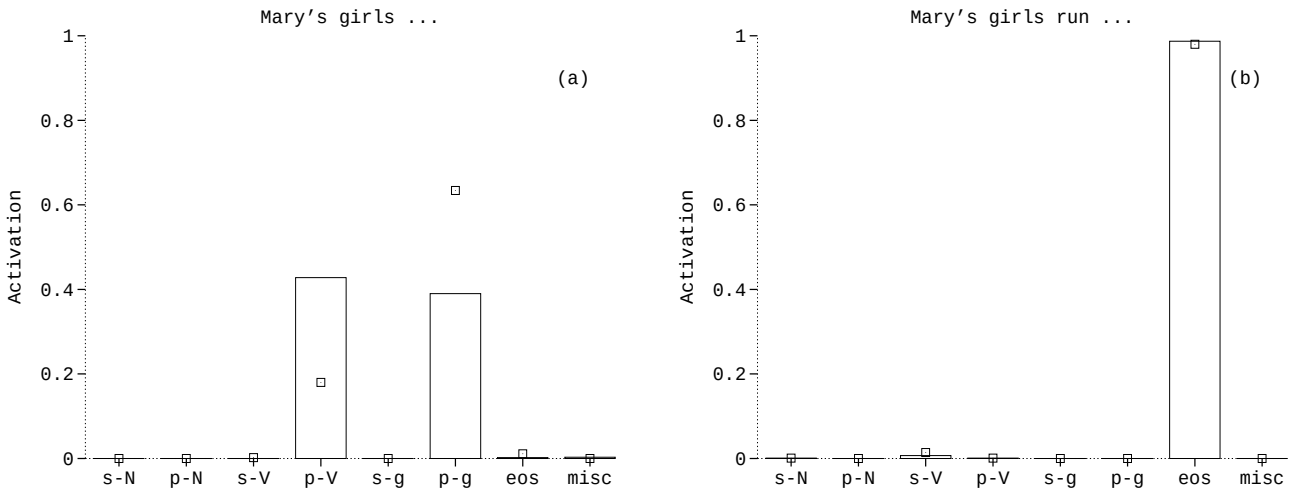


Figure 4.15. Network predictions after the last part of the test sentence *'Mary's girls run.'* (boxes) and in the control sentence *'Mary's cats run.'* (small squares).

shown after having received *'girls'* as input. The net was able to correctly activate the plural genitive marker in the test case even though it had never seen a genitive marker following *'girls'*. This indicates that the net is able to *strongly generalize* by predicting a known lexical item, the genitive marker, in a novel syntactic position, i.e., following *'girls'*. The activation of the plural genitive marker is not as high as the control, but it is nevertheless significant. Notice also that the prediction of a plural verb is stronger in the test case than in the control. Given the plural verb *'run'* in (b), the net is fully confident in its expectation of a end of sentence marker (both test and control). Importantly, strong generalization was also found in embedded sentences, such as, *'Mary's boys' girls run*, thus fulfilling the requirement of generalization to novel syntactic positions in both simple and embedded sentences. This positive finding becomes even more important, since Christiansen & Chater (1994) failed to obtain strong generalization in genitive contexts (although a minor activation of the plural genitive marker following *'girls'* did indicate that progress was possible).

Strong Generalization in NP Conjunctions

In contrast to Christiansen & Chater's problems in the genitive context, they reported a successful outcome of their testing of noun phrase conjunctions. This is also replicated below—albeit the results here evince a slightly higher degree of error than found in Christiansen & Chater. Figure 4.16 illustrates network behavior during the processing

warranted.

of the sentence *'Mary says that John and boy from town see'* in which *'boy'* occurs in the novel syntactic context of a noun phrase conjunction (contrasted with the control sentence *'Mary says that John and man from town see'*)[22].

The patterns of activation produced by the first four words have been depicted earlier. Given the context *'Mary says that John ...'*, the net predicts that the subsequent word must be either a singular verb, or a modification of *'John'* starting with a preposition, a conjunction, a singular genitive marker, or *'who'* (the latter in the `misc` group). The net also minimally activates the nouns which are not permitted here. In (b), the net is rightly confident that only a noun can come next[23]. Already in (c), we see the net's ability to strongly generalize in the NP conjunction case when it activates the plural verbs to match the conjoined NP *'John and boy'*. Recall that the net has only seen a singular verb following after *'boy'*. This means that the net has to 'overwrite' the statistical correlation between *'boy'* and single verbs in order to make the correct generalization that the NP conjunction takes a plural verb. Admittedly, the activation of the plural verbs are not as high as in the control sentence, but is still significant. Notice that despite the plural verb prediction, the net still expects that a singular genitive marker might be next (or, a preposition or *'who'*). The net exhibits a minor error by activating the singular verbs slightly (and does not reach the level of activation of *'and'* found in the control sentence). Since the input in (d) is the preposition *'from'*, the net predicts that the next word must be one of the nouns used with the prepositions. In (e) it is possible to detect two small errors in the net's predictions concerning the singular verbs and the end of sentence marker. This error is slightly more pronounced here than in the results presented in Christiansen & Chater. Nonetheless, the net still gets the plural verb agreement right across the prepositional phrase. As pointed out in Christiansen & Chater (1994), this is a considerable feat, since the net thus is able to strongly generalize *across several words*. In particular, it shows that the net is not simply predicting plural verbs on the basis of having seen an *'and'* two items before, but has learned the grammatical regularities subserving noun phrase conjunctions. Lastly, (f) demonstrates that not only is the net able to predict a correct end of sentence after *'Mary says that John and boy from town see ...'*, but it is also capable of predicting that *'see'* might take an optional direct object. As the net is also able to strongly generalize given NP conjunctions in simple sentences, such as, *'John and boy run'*, the net therefore fulfill the additional requirements for the ascription of strong generalization defined above.

[22] In Figure 4.16, `misc` contains only *'that'* and *'who'*.

[23] Again, it should be noted that neither *'boy'* nor *'boys'* receive anything but very minimal activation compared with the other nouns. I explain below why this should not be considered problematic.

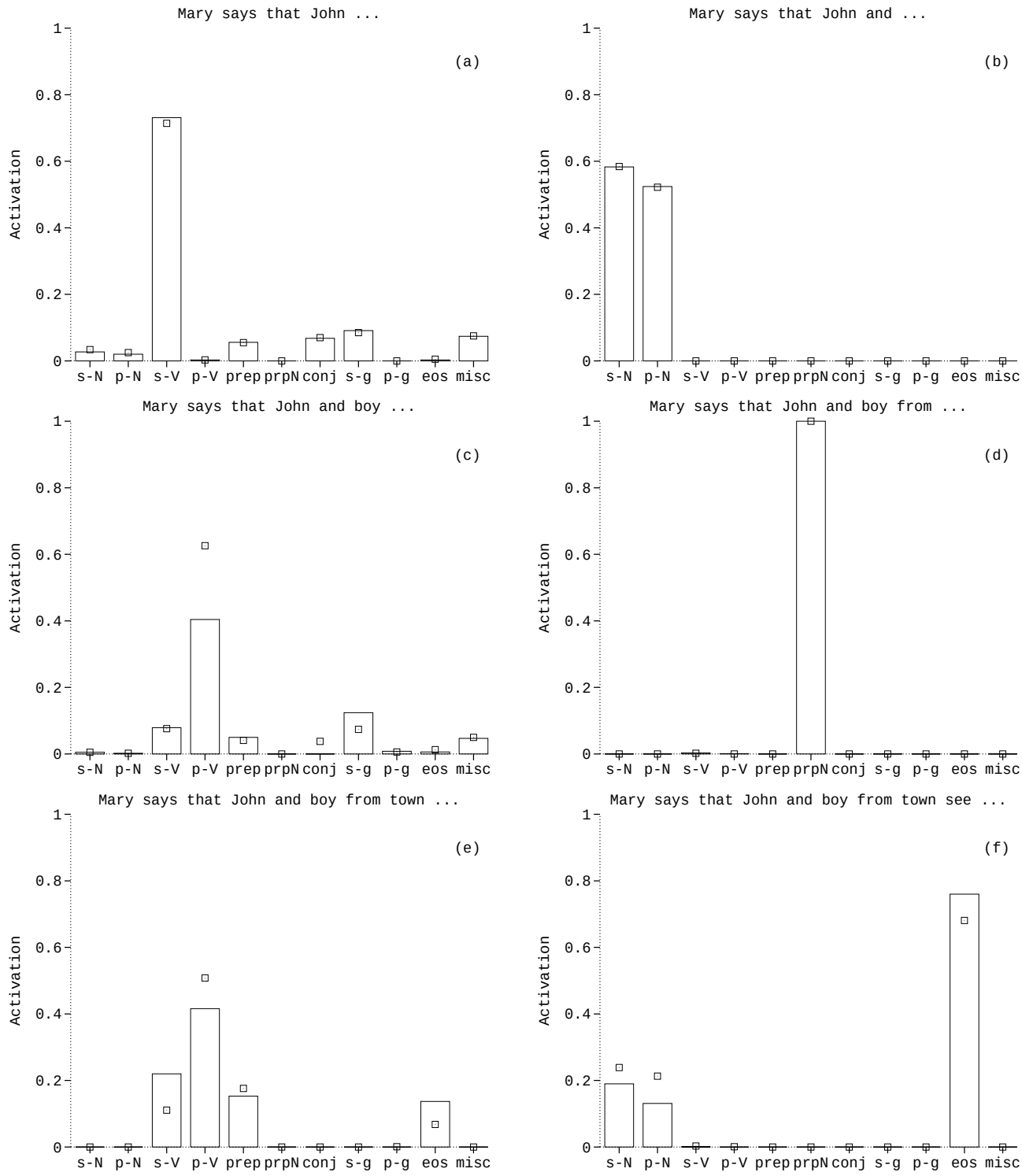


Figure 4.16. Network predictions after each word in the test sentence *'Mary says that John and boy from town see.'* (boxes) and in the control sentence *'Mary says that John and man from town eat.'* (small squares).

While describing network behavior during the processing of the test sentences I noted that it was not able to predict 'girl' and *'girls'* following a genitive marker and *'boy'* and *'boys'* following a conjunction. Hadley (1994b) raises doubts about whether the nets therefore can be said to exhibit genuine strong generalization. He suggests that they fail to do so: "it is difficult to see why the network's ability to predict *only previously encountered* nouns, at the crucial novel position, should in any way count as a success (at that position), given that we are testing for the network's generalization of *other* nouns into that position" (p. 13). Now, adults, and presumably children too, are able to predict subsequent words given a particular context (Grosjean, 1980). If Hadley's criticism is taken at face value, we would expect children to be able to predict newly learned words in novel syntactic positions. This seems highly unlikely. For example, if a child had never heard the word *'boy'* in a NP conjunction context, then she would arguably never predict *'boy'* in this position either (if somehow tested)—unless she had semantic or contextual information to tell her otherwise. This is exactly what is the case in the language learning studies that Hadley (1994a) refers to (e.g., Gropen, Pinker, Hollander, Goldberg & Wilson, 1989; Pinker, Lebeaux & Frost, 1987). In these studies children are taught nonsense words, such as, *'pilk'*, and primed to apply them in a novel syntactic context. It seems clear from the elaborate set-up in these experiments that this kind of generalization can only be obtained under conditions of strong semantic and contextual priming. These studies furthermore acknowledge that children generally appear to be rather conservative language learners. Of course, Hadley (1994b) is right to say that "children are capable of *spontaneously producing* sentences containing words in novel positions" (p. 13), but such occurrences are rare and, I submit, cued by semantic and/or contextual information. However, the networks used in Christiansen & Chater (1994) and here do not have such information available. The nets can therefore hardly be blamed for not predicting nouns in novel syntactic positions when no child, relying solely on syntactic information, is likely to be able to perform this task either.

Given these considerations, the ascription of strong generalization seems warranted in the test cases presented here. Whether these two instances of strong generalization are sufficient to endow the system with Hadley's notion of strong systematicity depends on whether four nouns out of a total number of ten nouns count as a "*significant fraction* of the vocabulary" (Hadley, 1994a, p. 7). Independent of the answer to this question, we may agree with Hadley that human language learners presumably are able to strongly generalize in a number of different syntactic contexts, more than reported here. Yet the net's ability to strongly generalize in both genitive and conjoined NP contexts suggests that this more widespread, human-like (strong) generalization may

not be beyond simple recurrent networks.

4.4 Discussion

In the current chapter, I have reported a number of simulation results extending those presented in chapter 3. In particular, the results here show that simple recurrent networks appear to have sufficient computational power to induce quite complex grammars involving both i-recursive structures and ni-recursive structures (the latter instantiated as either center-embedding or cross-dependency). Importantly, the nets exhibit the same kinds of limitations observed (or predicted to be found) in human performance on both ni- and i-recursive sentence constructions. Moreover, the nets also acquired an ability for strong generalization (at least, in the two test cases discussed above) comparable to what we may expect from human language learners. Simple recurrent networks therefore seem to be viable candidates as models of certain important aspects of human language acquisition.

One limitation of the simulations put forward in both this and the previous chapter is that they do not incorporate semantic or other non-syntactic information. Whether such a separation is warranted has been the matter of some controversy. Some researchers (e.g., Ferreira & Clifton, 1986; Perfetti, 1990; Rayner, Carlson & Frazier, 1983) have suggested that syntactic processing is autonomous from other kinds of language processing. Others again (e.g., Altmannn & Steedman, 1988; Crain & Steedman, 1985; Taraban & McClelland, 1990) have argued that semantic and contextual information indeed does affect syntactic processing. Although the debate is undecided, recent experimental evidence reported in, for example, Spivey-Knowlton & Tanenhaus (1994) does point towards the latter position. What is important in the present context is that we should, at least, be able to envisage how non-syntactic information may be incorporated in the currently purely syntactic account of language acquisition and processing. Fortunately, a simulation experiment conducted by Weckerly & Elman (1992) shows some indication of how this might be done. They trained a simple recurrent network on sentences with certain constraints on subject noun/verb combination. More specifically, certain animate verbs would only occur with animate subject nouns, whereas other varied freely between animate and inanimate subjects. When testing the net on center-embedded structures, a difference was observed in net performance depending on whether the verbs in the those sentences were biased towards the animacy/inanimacy of their subject nouns or neutral. As also reported in studies of human subjects tested on semantically biased and neutral center-embedded sentences (e.g., King & Just, 1991; Stolz, 1967), such bias clearly facilitates processing. Of course, the word co-occurrence

bias in the Weckerly & Elman's training set does not correspond to semantics proper. However, it does suggest that children may initially use statistical contingencies in language to bootstrap their further development of semantics (in a linguistic context—see also Finch & Chater, 1993, for a similar statistically motivated view). Pointers in the direction of implementing something closer to contextual processing may be found in St. John & McClelland's (1990) recurrent network model of the learning and application of contextual constraints in sentence processing which incorporates incremental interpretation. Although extending the simulation results presented in here, and in chapter 3, is not trivial, the results reported in Weckerly & Elman (1992) and St. John & McClelland (1990) indicate, at least, that such an extension is not impossible in principle.

Another possible limitation concerns the simulation results on strong generalization presented in the latter part of this chapter. It is often noted that connectionist and other bottom-up statistical models of language learning will not be able to scale up to solve human language acquisition because of arguments pertaining to the purported poverty of the stimulus. I address these arguments at length in the next chapter. Here it suffices to say that there are some evidence that models employing simple statistical analysis may be able to attain strong generalization. Christiansen & Chater (1994) mention that when Redington, Chater & Finch (1993) applied a method of distributional statistics (see also Finch & Chater, 1992, 1993) to a corpus of child directed speech (the CHILDES corpus collected by MacWhinney & Snow, 1985), they found that the syntactic category of a nonsense word could be derived from a single occurrence of that word in the training corpus. This indicates that strong generalization may be learnable through bottom-up statistical analysis—even on a scale comparable with that of a child learning her first language. In this context, it is also important to note that achieving strong generalization is not only a problem for connectionist models of the learning of linguistic structure. As pointed out by Christiansen & Chater (1994), most symbolic models cannot be ascribed strong generalization since they in most case are spoon-fed the lexical categories of words via syntactic tagging. The question of strong generalization is therefore just as pressing for symbolic approaches as for connectionist approaches to language acquisition. The results presented in this chapter suggest that connectionist models may be closer to solving this problem than their symbolic counterparts.

Up until now, I have focused on establishing that connectionist models have sufficient computational power and generalization capability to serve as models for natural language learning and processing. Next, in chapter 5, I outline a theory of how our

language ability may have evolved to its present state (as modeled, in part, here and in the previous chapter). Learning plays an important role in this account of the origin and evolution of language, and the observations made in the present chapter regarding the incremental memory training regime provide partial support for an explanation of how children may overcome the apparent poverty of the stimulus.

Chapter 5

The Evolution and Acquisition of Language

When studying natural language one cannot help being filled with awe over its intricate, yet highly regularized, complexity. Moreover, the speed at which children acquire this formidable means of communication must strike even the most casual observer with amazement. How do children accomplish this enormous feat? Do they *learn* to speak their native tongue? Or, does their language ability gradually unfold according to a genetic blueprint (in much the same way that a chicken grows a wing)? Thus, the question is whether there is sufficient environmental information available to the child to make language learnable (given *general* non-linguistic constraints on learning), or, is it necessary to presuppose the existence of *specific* linguistic constraints in order to account for language acquisition?

For a child to acquire language it is clear that whatever mechanisms participate in this task, they would have to be biased in some way towards the learning of language—or, at least, towards the learning of sequential and hierarchical structure. Otherwise, we would be able to teach computers to speak simply by talking to them whilst they record our speech. In other words, that there must be internal constraints on the acquisition of language is hardly controversial, but the nature and extent of these constraints is the focus of much debate. For a period spanning three decades Chomsky (1965, 1972, 1975, 1976, 1977, 1980, 1986, 1988, 1993) has argued forcefully that a substantial innate endowment of language specific knowledge is necessary in order to provide sufficient constraints on language acquisition. These constraints form a *'Universal Grammar'* (UG); that is, an innate database consisting of a collection of universal grammatical principles that hold across all human languages. In this framework, all that language 'learning' amounts to is the setting of a number of parameters in UG according to

the specifics of the particular language being learned. The staunchest proponents of this view even go as far as to claim that "doubting that there are language-specific, innate computational capacities today is a bit like being still dubious about the very existence of molecules, in spite of the awesome progress of molecular biology" (Piattelli-Palmarini, 1994: p. 335).

Given this view of language acquisition, a question naturally arises concerning the evolution of such an elaborate and highly specialized innate structure. It is often noted that humans appear to be the only species in which nature has bestowed language (at least, in its present complexity). But how did the human language ability come about in the first place? What kind of phylogenetic explanation might be found for this uniquely human capacity? The proponents of UG are generally divided into two camps when addressing the issue of language evolution. One camp (e.g., Chomsky, 1988; Piattelli-Palmarini, 1989) has suggested that natural selection only played a minor role in the emergence of language in humans. On this account, UG is a product of *exaptation*; that is, it is hypothesized that it might have arisen as a by-product of increased brain size following evolutionary pressures driven by other functions than language, or, perhaps, as a consequence of random mutations. The other camp (e.g., Bloom, 1994; Corballis, 1992, 1994; Greenfield, 1991; Hurford, 1991; Pinker, 1994; Pinker & Bloom, 1990) emphasizes a gradual evolution of the human language faculty through *natural selection*. In this picture, it is assumed that having a language confers added reproductive fitness on humans and that this, in turn, leads to a selective pressure towards increasingly more complex grammars.

Both accounts of the evolution of language do not leave much room for learning. The proponents of neo-Darwinian evolution do, however, leave a little elbow room for learning inasmuch as evolution through natural selection—being essentially a hill-climbing process (Hinton & Nowlan, 1987; Maynard-Smith, 1987; Pinker & Bloom, 1990)—can be construed as a kind of (non-Larmarckian) learning process in which a particular species searches the evolutionary problem space for good genotypes (albeit that the species has no 'memory' of its previous searchpath since individuals with poorly adapted genotypes tend not to live to tell the tale). The exaptationists seek to abolish the term "learning" altogether, suggesting "that we would gain in clarity if the *scientific* use of the term were simply discontinued" (Piattelli-Palmarini, 1989: p. 2; see also Chomsky, 1980, for a similar view). Given that learning generally plays a fundamental role in most connectionist theories, and in the work presented in the previous chapters in particular, such eschewal of the concept of (non-trivial) learning

within cognitive science would have a devastating impact on connectionist research[1]. In this chapter, I therefore sketch a theory in which considerations regarding learning and processing, rather than innate linguistic knowledge, provide the explanation of language development and evolution. This involves a reappraisal of the poverty of the stimulus arguments typically presented in favor of an innate UG, suggesting an alternative account of the psycholinguistic data based on a dynamic perspective on the co-evolution of language and the human mechanism underlying both learning and processing of linguistic structure. If this account is correct, then it is a mistake to bury a scientific term which is still very much alive and kicking. Indeed, I predict that the concept of learning is destined to play an important part in future research into language and other parts of cognition.

The content of this chapter is as follows: First I examine the main exaptationist and adaptationist perspectives on the evolution of language. An alternative view is proposed in which language is treated as an organism that is forced to adapt to the idiosyncrasies of its human hosts. To exemplify the kind of explanation of linguistic phenomena that this approach may offer, I discuss subjacency as a classic example of an arbitrary language universal. Next, in section 2, I address the issues concerning the origin of language. It is argued that sequential learning is the basis of our language ability. The latter is hypothesized as having started as a manual language which gradually evolved into a predominately vocal language following bipedalism and changes in the human vocal tract. I furthermore contend that such a learning based language capability could not have become innate following subsequent evolution. An account of linguistic change is then put forward, pointing to increases in vocabulary size as the key factor in bringing about morphological and syntactic change. The emphasis on learning in the evolutionary scenario is continued in section 3 which discusses the acquisition of language. Innate, but not language-specific, maturational constraints are advocated to determine the acquisition process, providing a plausible explanation of the critical period of language learning. The argument from the poverty of the stimulus is re-appraised, and it is argued that language may be learnable without the help of the massive endowment of innate linguistic knowledge presupposed by the proponents of UG. Finally, two possible objections to the present theory are debated.

[1]It should be noted that connectionism, in principle, might be compatible with some kinds of nativism (cf. e.g., Clark, 1993; Ramsey & Stich, 1991), but it seems clear that the *spirit* of connectionism is incompatible with the strong innateness hypothesis espoused by, for example, Crain (1991), Chomsky (1980, 1986, 1993) and Piattelli-Palmarini (1989, 1994).

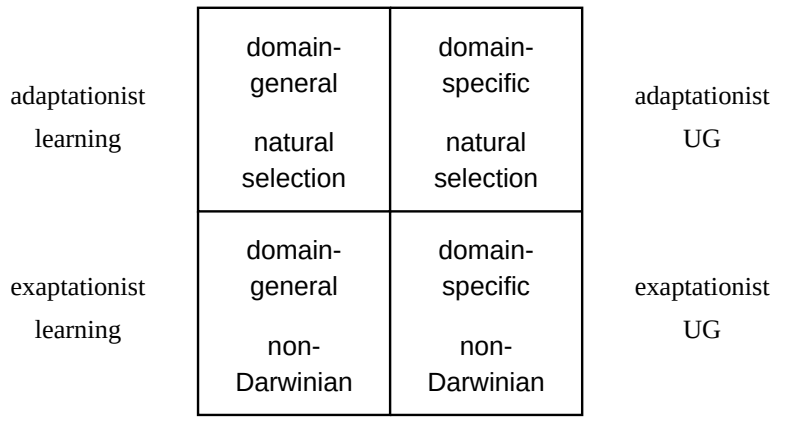


Figure 5.1. Schematic representation of four positions concerning explanations of the acquisition (domain-general vs. domain-specific) and evolution (natural selection vs. non-Darwinian) of language.

5.1 Language: Organ, Instinct, or Nonobligate Symbiant?

Ultimately, language has to be tied to the phylogeny and ontogeny of human biology. In an attempt to characterize the biological underpinnings of language, Chomsky (1965, 1986, 1988) has advocated that language should be viewed as one amongst many *'mental organs'* which "develop in specific ways each in accordance with the genetic program, much as bodily organs develop; and that multipurpose learning strategies are no more likely to exist than general principles of 'growth of organs' that account for the shape, structure, and function of the kidney, the liver, the heart, the visual system, and so forth" (Chomsky, 1980: P. 245). More recently, Pinker (1994) has argued that language is better construed as an *instinct* because "it conveys the idea that people know how to talk in more or less the sense that spiders know how to spin webs" (p. 18). Both terms carry much the same nativist commitments on their sleeves, indicating that only highly domain-specific 'trigger' learning can take place. Yet, the two positions diverge substantially on what role natural selection is meant to play in the evolution of UG.

The general relationship between the two perspectives is illustrated in Figure 5.1, where the *'exaptationist UG'*[2] position is taken by Chomsky (1972, 1982, 1988, 1993)

[2]It should be noted that I here follow Piattelli-Palmarini (1989) in using 'exaptation' as an umbrella term for recent nonadaptationist mechanisms for evolution, such as, *genetic hitch-hiking* (Maynard-Smith, 1978), that is, a mechanism by which non-selected genes might "catch a ride" with another gene that *was* selected for, if they are in close proximity to the selected gene along a chromosome; *spandrels* (Gould & Lewontin, 1979), i.e., architectural by-products with no previous function, but which come to serve some novel function (by analogy to the mosaics on the triangular spaces formed at the intersection of the arches of the dome in the San Marco basilica in Venice); and, *exaptation* proper

and Piattelli-Palmarini (1989, 1994) whereas the (seemingly more popular) *'adaptationist UG'* view counts Bloom (1994), Corballis (1992, 1994), Hurford (1991), Pinker (1994), Pinker & Bloom (1990) amongst its proponents. A third perspective emphasizing both natural selection and genuine (domain-general) learning in the phylogeny and ontogeny of language—here named *'adaptationist learning'*—is held by, e.g., Bates, Thal & Marchman (1989), Bates & Elman (1993) and Elman (1993). The logical structure of Figure 5.1 suggests a fourth way of looking at the issues involved in the evolution and acquisition of language: the *'exaptationist learning'* viewpoint[3]. To my knowledge only one person comes close to having this point of view. In arguing for his view of language as a spandrel, Gould (1993) acknowledges that he "can't prove that language was not the selected basis of increasing brain size, but the universals of language are so different from anything else in nature, and so quirky in their structure, that origin as a side consequence of the brain's enhanced capacity, rather than as simple advance in continuity from ancestral grunts and gestures, seems indicated" (p. 321). Elsewhere, Gould (1979) has made comments which could be interpreted (as they have, indeed, by Pinker & Bloom, 1990) as suggesting that the increased brain size produced a multipurpose learning device that can acquire not only language, but also many other cognitive abilities.

5.1.1 The Exaptationist View

The other exaptationist viewpoint—though certainly not the most popular account of language development and evolution—has its prominent advocates. For instance, Chomsky (1972, 1982, 1988) has for more than two decades expressed strong doubts about neo-Darwinian explanations of language evolution and has recently been joined by Piattelli-Palmarini (1989). This skepticism does not merely concern adaptationist accounts of language origin, but the selectionist theory of evolution as a whole:

> What Darwin achieved is of extraordinary importance, but there's virtually nothing of a theory there ...when you try to account for why particular organs develop, or species, and so on, all you can do is wave your hand ...To move to more far reaching explanation, you're going to have to find

(Gould & Vrba, 1982), that is, when something that was originally adapted to serve a particular function is put to use to serve a novel function.

[3]I would like to stress that I do not intend the four positions represented in Figure 5.1 to exhaust all possible perspectives on the evolution and development of language. Indeed, my own view, as we shall see, falls outside this schematization. I have included this figure in order to provide a clear schematic representation of the relations between the three viewpoints most often found in the recent literature focusing on the issues at hand. The possibility of the exaptationist learning position follows logically from the figure, but does not seem to have much support.

> something about the space of physical possibility within which selection operates. That space might be extremely narrow. For example, it might be so narrow that under the particular conditions of human evolution, there's one possibility for something with 10^{11} neurons packed into something the size of a basketball: namely, a brain that has these computational properties. (Chomsky, 1993: p. 83)

Chomsky is careful to add that he is not proposing such an evolutionary picture. Indeed, he does not commit himself to any particular view of evolution[4]. Chomsky does, however, show a strong inclination towards an exaptationist framework, rather than an adaptationist one (also cf. Pinker & Bloom, 1990). Piattelli-Palmarini (1989), on the other hand, demonstrates an even stronger commitment to exaptationism, finding adaptationist explanations much too weak: "Adaptive constraints are typically insufficient to discriminate between real cases and an infinity of alternative, incompatible mechanisms and traits which, although abstractly compatible with the survival of a given species, are demonstrably absent" (p. 19).

The exaptationist positions from Figure 5.1 both rely on the complexity and intricacy of the putative UG as the premise for their arguments against adaptationist explanations of language evolution. UG appears to be so unique in terms of structure and properties, that it is unlikely that it could be a product of a process of natural selection amongst random mutations, or so the argument goes. Instead, Chomsky (1988) has suggested that, perhaps, the property of 'discrete infinity' (recursion) arose as a consequence of a mutation in some protohuman being, making language possible. Elsewhere, Chomsky (1972, 1982, 1993) proposes that the language organ might be a by-product of having a brain of a particular size and structural complexity following certain (yet unknown) properties of physical mechanisms. This view of language as spandrel has received further support from Gould (1993). Moreover, in reference to the arbitrariness of the specific set of principles and parameters that characterizes UG, Piattelli-Palmarini (1989) has noted that "adaptationism cannot even begin to explain why the natural languages that we can acquire and use possess these central features and not very different ones" (p. 24).

5.1.2 The Adaptationist Perspective

Recently, Pinker & Bloom (1990) have forcefully defended the adaptationist perspective on language from the exaptationist attacks (briefly summarized above). They, too, adopt UG as a premise for their arguments (thus, espousing the adaptationist UG

[4]Elsewhere, for instance, Chomsky says in an often quoted passage that "language must surely confer enormous selective advantages" (1980, p. 239).

position in Figure 5.1). But in contrast to Chomsky, Gould and Piattelli-Palmarini, Pinker & Bloom find that the complex and intricate structure of UG bears evidence of design, indicating its adaptation as an efficient means of communication. More generally, they argue that "*natural selection is the only scientific explanation of adaptive complexity.* 'Adaptive complexity' describes any system composed of many interacting parts where the details of the parts' structure and arrangement suggest design to fulfill some function" (p. 709; their emphasis). As another example of adaptive complexity, they refer to the vertebrate visual system. This system consists of many parts that have to work together to create vision; starting with the refracting cornea and the illumination sensitive pupil (regulating the amount of incoming light), which allow light to impinge on the retina through a variable focus lens. The image is then transmitted via the optical nerves to the visual cortex where specialized neural structures respond to various parts of the input patterns, such as, edges, color, motion, and so on. Moreover, the eyes are equipped with muscles that ensure coordination between the two eyes as well as visual stability (for instance, to maintain fixation on a given point in visual space when the head is moving). Pinker & Bloom argue that such an arrangement of matter has an extremely low probability of occurring by chance because of the complex interactions amongst its highly specialized parts. Consequently, they find that it would amount to something close to miracle for the vertebrate visual system to emerge as a product of random mutation (e.g., via genetic hitch-hiking), some (yet unknown) laws concerning the possible biological arrangements of matter, recycling of a structure adapted for another function (exaptation proper), or as the byproduct of architectural constraints of other unrelated structures (a spandrel).

Given that language couched in terms of UG appears to show a degree of complexity equal to that of the vertebrate visual system, Pinker & Bloom conclude that it is highly improbable that language is the product of some nonadaptationist process. In particular, they contend (pace Chomsky, Gould and Piattelli-Palmarini) that language cannot be an *unmodified* spandrel. Although unmodified spandrels appear to exist (e.g., as when wading birds use their wings primarily to block out the reflections of the sun while fishing), each such exaptation is merely one amongst many crude solutions to a simple engineering task, and is therefore not likely to play a dominant role in evolution of more complex capacities. On the other hand, *modified* spandrels—that is, cases where spandrels are redesigned to serve new functions—play a much more important role in evolution because they provide starting points for adaptive complexity. As an example, Pinker & Bloom mention the evolution of insect wings which involved the redesigning

of structures that were originally evolved for the purpose of thermal exchange[5]. Thus, exaptationist processes appear to provide the raw material for natural selection, which in turn, never has a clean slate to start with, but must always work as tinkerer with whatever building materials are at hand. Importantly, it is adaptation that shapes and finetunes the structures underlying complex functions.

Pinker & Bloom acknowledge that language might be a modified spandrel that evolved from general cognitive processes not specific to language. In particular, they speculate that "the multiplicity of human languages is in part a consequence of learning mechanisms existing prior to (or at least independent of) the mechanisms specifically dedicated to language" (1990: p. 723). Such learning mechanisms subserving early language would then have been shaped by natural selection into our presentday language ability via natural selection. On this view, what originally had to be learned gradually became innate following the *'Baldwin effect'* (Baldwin, 1896; Hinton & Nowlan, 1987; Maynard-Smith, 1987), thus creating the UG that Pinker & Bloom take to underlie human language. The basic idea behind the Baldwin effect is that when a species is faced with the learning of an adaptive task, certain initial settings of the learning mechanism are better than others. Individuals endowed with good starting configurations are likely to learn faster, which, in turn, confers added reproductive fitness onto them. These individuals should therefore proliferate, creating offspring with equally good starting points for learning (genetic drift aside). This process repeats itself, producing better and better initial settings for learning, until the task becomes more or less innately specified. At this point learning might merely consist in the triggering of a few switches in an otherwise innate database. The Baldwin effect thus provides a way in which learning can guide evolution in a nonLamarkian way, and is used by Pinker & Bloom to explain how the innate UG might have evolved from language independent learning mechanisms (for a criticism of this view, see section 5.2.3).

The Arbitrariness of Linguistic Universals

Having thus defended the adaptationist UG position against the onslaught of the exaptationists, Pinker & Bloom still have to explain why the principles of UG are essentially arbitrary (as pointed out by Piattelli-Palmarini, 1989). They address this issue by suggesting that the constraints imposed by UG function as communicative protocols. As such, the specific nature of these standards does not matter as long as everyone (within

[5]It is amusing to note that if the Disney character "Dumbo" was a real living organism, then it would seem to have followed the same path to flight as the insects; that is, a transformation of its ears (as heat exchangers) into structures subserving flight. However, this is not to say that there might not be laws of nature that generally prevent elephants from ever becoming airborne by their own power.

a given speech community) adopts the same set of standards. The arbitrariness of the principles of UG can therefore be seen to parallel the arbitrariness of technical communication protocols, such as, those used in communication between computers. When using a modem it is important to use the right protocol; for instance, odd parity, handshake on, 7 bit, etc. There is no particular reason for having a protocol with exactly these settings, other combinations would be just as effective. What is important, however, is that the two computers that are to communicate with each other adopt the *same* protocol—otherwise, communication will not be possible at all. So, when it comes to the specifics of UG, Pinker & Bloom suggest that "in the evolution of the language faculty, many 'arbitrary' constraints may have been selected simply because they defined parts of a standardized communicative code in the brains of some critical mass of speakers" (1990: p. 718)[6].

Pinker & Bloom's (1990) elaborate account of the evolution of language still leaves some questions to be answered; an important one being: How did language get to have the structure that it has? The adaptationist UG view leaves the origin of language structure mostly unexplained, characterized as a collection of arbitrary communication standards. For example, Pinker & Bloom write that

> ...many aspects of grammar cannot be reduced to being the optimal solution to a communicative problem; rather, human grammar has a universal idiosyncratic logic of its own ...Evolution has had a wide variety of equivalent communicative standards to choose from; there is no reason for it to have favored the class of languages that includes Apache and Yiddish, but not Old High Martian or Early Vulcan ...Whatever rationales may have influenced these choices are buried in history.(p.719)

Pinker & Bloom further suggest that these idiosyncrasies are in part culturally determined, but this suggestion just pushes the question back one level: How did they evolve in the cultural domain? Thus, it seems to be something of a mystery that we only learn the human languages (with their arbitrary idiosyncrasies) given that they comprise a mere fraction of the total set of *theoretically* possible languages.

5.1.3 Language as an Organism

If we, however, invert the perspective on language evolution—recognizing that language has evolved to fit the human learning and processing mechanism—then the mystery

[6]In addition, Pinker & Bloom (1990) point out that it is often the case that natural selection has several (equally adaptive) alternatives to choose from to carry out a given function (for example, both the invertebrate and the vertebrate eye support vision despite having significant architectural differences.

can be unraveled; and we might, furthermore, understand how language got to have its apparently "idiosyncratic" structure. Instead of saying that humans can only learn a small subset of a huge set of possible languages, we must refocus by observing that *natural* languages exist only because humans can produce, learn and process them. In this connection, it is useful to construe language as an *organism*, adapted through natural selection to fit a particular ecological niche: the human brain. Darwin (1900) was one of the first to recognize this as is evident from the following quote:

> The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel ...We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation. The manner in which certain letters or sounds change when other change is very like correlated growth ...Languages, like organic beings, can be classed in groups under groups; and they can be classed either naturally, according to descent, or artificially by other characters. Dominant languages and dialects spread widely, and lead to the gradual extinction of other tongues. A language, like a species, when once extinct, never ...reappears ...A struggle for life is constantly going on among the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand ...The survival and preservation of certain favored words in the struggle for existence is natural selection. (p. 106)

In this sense, natural language is akin to an organism whose evolution has been constrained by the properties of human learning and processing mechanisms. It is therefore not surprising that we, after all, are so good at acquiring language. Language is closely tailored for human learning, rather than the other way round (as suggested by Pinker & Bloom, 1990: p. 712). In addition, it is also worth noting that the human language learning mechanism is not static (as we shall see in section 5.3). It undergoes significant changes during the period of (optimal) language acquisition. In the present picture, language evolution and development are tied strongly together (pace e.g., Chomsky, 1988, 1993; Piattelli-Palmarini, 1989; Pinker & Bloom, 1990; Pinker, 1994). Only by studying both in unison can we begin to understand how natural language came to be the way we experience it today.

Presentday natural language is not due to exaptation (of almost magical proportions) as suggested by Piattelli-Palmarini (1989), nor is its many universal features essentially arbitrary as implied by Pinker & Bloom's (1990) account. Rather, I contend that language is the product of an evolutionary process in which language had to adapt to the human learning mechanism (with all its developmental idiosyncrasies)

in order to 'survive'[7]. This is not to say that having a language does not confer selective advantage onto humans. It is clear that humans with a superior language ability are likely to have a selective advantage over other humans (and other organisms) with lesser communication powers. This is an uncontroversial point, forming the basic premise of many evolutionary theories of language origin (save Chomsky, 1988; Gould, 1993; Piattelli-Palmarini, 1989, 1994).

What is often not appreciated is that the selective forces working on language to fit humans is significantly stronger than the selective pressure on humans to be able to use language. In the case of the former, a language can *only* survive if it is learnable and processable by humans. On the other hand, adaptation towards language use is merely *one out of many* selective pressures working on humans (such as, for example, being able to avoid predators and find food). Whereas humans can survive without language, the opposite is not the case (at least, not as far as human languages—the focus of linguistics—is concerned). Thus, language is more likely to have adapted itself to its human hosts than the other way round. Languages that are hard for humans to learn simply die out, or, more likely, do not come into existence at all. Following Darwin, I propose to view natural language as a kind of a beneficial parasite—i.e., a *nonobligate symbiant*—that confers some selective advantage onto its human hosts without whom it cannot survive. Consequently, the rate of linguistic change is far greater than the rate of biological change. Whereas it takes about 10,000 years for a language to change into a completely different "species" of language (e.g., from protolanguage to presentday language, Kiparsky, 1976), it took our remote ancestors something in the neighborhood of 250,000 years to evolve from the archaic form of *Homo sapiens* into the anatomically modern form we have today, *Homo sapiens sapiens* (cf. data in Corballis, 1992). The fact that children are so successful at language learning is therefore more appropriately explained as a product of natural selection of linguistic structures, rather than natural selection of biological structures, such as UG.

Returning to the universal principles of UG and their supposedly arbitrary nature, it is clear that they *are* arbitrary from a linguistic point of view. That is, given a

[7]In this connection, it is interesting to note that a group of linguists recently also adopted the view of language as a kind of organism (although I presume that they would not agree with the conclusions that I draw here from this perspective). In a collection of papers Hale *et al* (1992) express their worries about the rapidly increasing number of endangered languages; that is, languages that have disappeared, or are about to disappear, from the face of the earth. In one of the papers, Michael Kraus suggests that "language endangerment is significantly comparable to—and related to—endangerment of biological species in the natural world" (p. 4). He goes on to warn us that "the coming century will see either the death or doom of 90% of mankind's languages" (p. 7). This kind of language extinction is, however, not a product of natural selection (with respect to the human brain), but a product of a general pressure towards cultural homogeneity (cf. Kraus).

strictly linguistic perspective on language these constraints would appear to be arbitrary, since we can imagine a multitude of alternative, and equally adaptive, constraints on linguistic form. For instance, Piattelli-Palmarini's (1989) contends that there are no (linguistic) reasons not to form yes-no questions by reversing the word order of a sentence instead of the normal inversion of subject and auxiliary. However, on the present account linguistic universals are no longer arbitrary. Rather, they are determined by the properties of the human learning and processing mechanisms that underlie our language capacity. This is why we do not reverse the word order to form yes-no questions; it would put too heavy a load on memory to store a whole sentence in order to be able to reverse it[8]. In effect, language universals are by-products of processing and learning under certain limitations on memory, attention, etc. (and this, as we shall see in section 5.3, surprisingly makes language easier to learn). Consequently, there are good reasons why we are able to speak Apache and Yiddish, but not Old High Martian or Early Vulcan (pace Pinker & Bloom, 1990): the latter, non-human languages did not evolve because they simply do not fit the human learning and processing mechanisms. However, if we imagine that these brain mechanisms had followed a different evolutionary path, then we might have both Old High Martian and Early Vulcan amongst presentday human languages, but not Apache and Yiddish. Whereas the make-up of the human language machinery is arbitrary (it is, at least, conceivable that it could have been different), the structure of the human languages are not, since they are evolutionarily customized to fit human learning and processing capabilities. In short, my view amounts to the claim that most—if not all—linguistic universals will turn out to be terminological artifacts referring to mere side-effects of the processing and learning of language in humans[9].

Subjacency: A Classic Example of Arbitrariness

Since the subjacency principle, according to Pinker & Bloom (1990: p. 717), "is a classic example of an arbitrary constraint", it is well-suited as a demonstration of what

[8]Besides, this also presupposes that transformations must underlie the construction of yes-no questions—a point which *is not* an established truth (as we shall see later).

[9]Recently, Chomsky (1993) has expressed a somewhat similar view when outlining his new 'minimalist' program: "Grammatical constructions such as relative clause, passive, verbal phrase, and so on, appear to be taxonomic artifacts, like 'terrestrial mammal' or 'household pet'; the array of phenomena derive from the interaction of principles of much greater generality ... these principles may themselves be epiphenomenal, their consequences reducing to more general and abstract properties of the computational system, properties that have a kind of 'least effort' flavor" (p. 51). However, it is clear that Chomsky takes a rather different position on most other issues involved in the evolution and development of language.

my position implies[10]. The grammatical theory underlying Chomskyan UG (that is, Government and Binding (GB), e.g., Chomsky, 1981) involves an essential distinction between S-structures and D-structures (formerly, e.g., Chomsky, 1965, called "surface structures" and "deep structures", respectively), where the former is derived transformationally from the latter via "move-α", a general procedure for the movement of constituents. In order to restrict this powerful movement procedure, so that the transformed sentence structures correspond to what is acceptable to a normal speaker, a number of constraints are imposed on such transformations—subjacency being one of these. Thus, the subjacency principle involves certain general restrictions on the movements of constituents during transformation (the exact details are irrelevant here). Consider the following:

(1) *Who does Anita believe Betty bit* (gap)?
(2) $[_{\overline{S}_1}$ $comp_1$ *Anita believes* $[_{\overline{S}_2}$ $comp_2$ *Betty bit who*]]

In GB, the gap at the end of the S-structure in (1) is a result of two consecutive movements of *'who'* from its position at the end of the D-structure in (2). These movements are shown in the syntactic tree in Figure 5.2. Now, consider the following ungrammatical (S-structure) sentence (3) and its underlying D-structure (4):

(3) **Who does Anita believe the story that Betty bit* (gap)?
(4) $[_{\overline{S}_1}$ $comp_1$ *Anita believes* $[_{NP}$ *the story* $[_{\overline{S}_2}$ $comp_2$ *Betty bit who*]]]

As before, the gap in (3) is due to the movement of *'who'* from it tail position in (4). In order to explain the ungrammaticality of (3) compared with (1) the subjacency principle—formalized in Figure 5.3—is invoked. By comparing Figure 5.2 and 5.4 in the light of 5.3, we can now see why (1) is rendered grammatical within the GB framework and (3) is not. Figure 5.4 shows that the second movement of *'who'* in (3) is illegal because it results in a movement across two bounding nodes (NP and $\overline{S}_2$). This is prohibited by the subjacency principle. As such, the subjacency principle seems to lend itself easily to an explanation in terms of memory limitations (also cf. Elman, 1992). Indeed, Berwick & Weinberg (1984) have provided a functional explanation of subjacency. In their framework, the subjacency principle emerges from

[10]I have included this somewhat elaborate example to provide some linguistic 'weight' to the present theory and to make the discussion of linguistic universals more concrete. The example is furthermore presented as a preliminary template for the explanation of other linguistic universals as well as indicating how it may cut across the particulars of different linguistic frameworks (hence the GPSG part of the example).

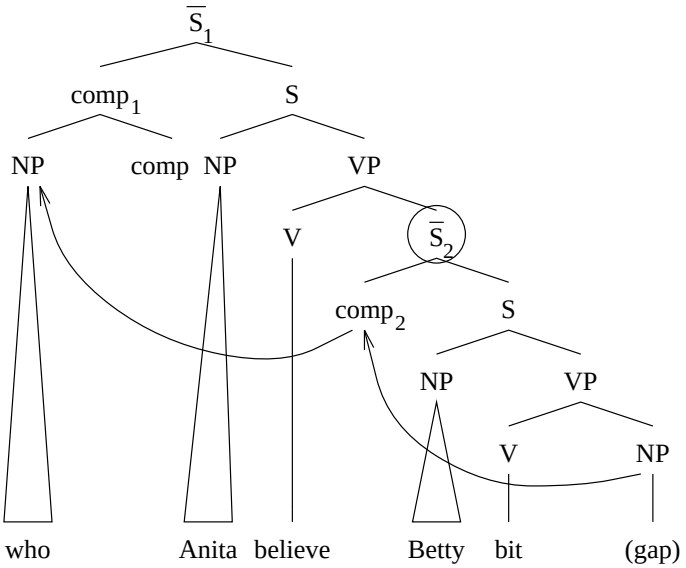


Figure 5.2. A Chomskyan (GB) transformational derivation of of the S-structure *'Who does Anita believe Betty bit'* from the D-structure *'Anita believes Betty bit who'* with the arrows illustrating the cyclic movements of *'who'*.

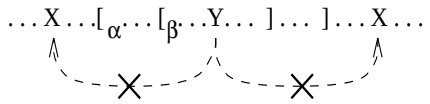


Figure 5.3. A formalization of the subjacency principle: no transformation may move a constituent from position Y to either of the X positions; that is, no single movement is allowed across more than one boundary node (where α, β—i.e., the bounding nodes—are either NP or $\overline{S}$).

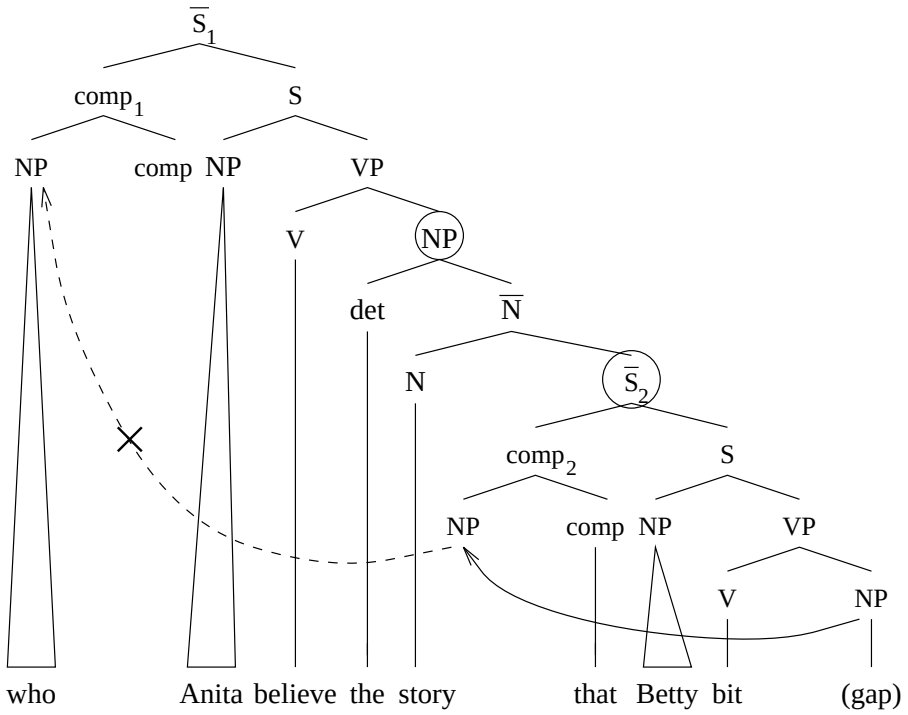


Figure 5.4. A Chomskyan (GB) derivation of of the ungrammatical S-structure *'Who does Anita believe the story that Betty bit'* from the D-structure *'Anita believes the story that Betty bit who'* with the plain arrow showing the first, legal movement of *'who'* and the dashed arrow the second, illegal movement across two bounding nodes (NP and $\overline{S}_2$).

the limitations on stack depth, when a GB grammar is implemented in a deterministic parsing system. Since we therefore reasonably can construe subjacency simply as a constraint on processing (or performance cf. chapter 2), it can no longer be considered to be an arbitrary linguistic phenomenon (as suggested by Pinker & Bloom, 1990), but must instead be conceived as a nonarbitrary byproduct of limited human processing abilities.

At this point it is furthermore illuminating to recall that the putative principles of UG are *not* established, scientific facts (even though Piattelli-Palmarini, 1994, and others would have us believe so). The GB framework underlying UG (as expounded by, e.g., Chomsky, 1957, 1965, 1976, 1981, 1986; Crain, 1991) is merely one amongst many linguistic theories—albeit perhaps the most dominant one. Many alternative theories exist in the realm of linguistics, such as, for example, Categorial Grammar (Steedman, 1987), Cognitive Grammar (Langacker, 1987), Dependency Grammar (Hudson, 1990), Lexical Functional Grammar (Kaplan & Bresnan, 1982), and Generalized Phrase Structure Grammar (GPSG: Gazdar, Klein, Pullum & Sag, 1985). Hence, it is possible to

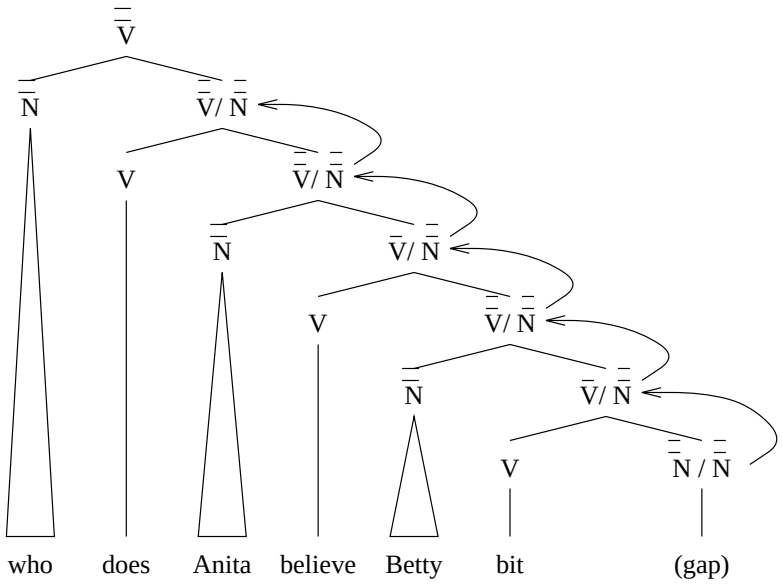


Figure 5.5. A syntactic tree derived via GPSG for the grammatical sentence *'Who does Anita believe Betty bit'* with the arrows showing how the gap (i.e., that a $\overline{\overline{N}}$ is needed) is percolated up the tree and discharged at the root.

adopt a number of different analyses of the linguistic data, some of which do not involve transformations, and therefore explain the ungrammaticality of sentences, such as (3), without reference to subjacency. For example, in GPSG (1) would result in the syntactic tree displayed in Figure 5.5. The gap (that is, the missing $\overline{\overline{N}}$) is simply percolated up the tree until it can can be discharged at the root. Notice that in GPSG syntactic trees are basically a collection of local trees of depth one (the exact nature of GPSG is irrelevant for the purpose of the present comparison with GB). A grammatical principle, such as subjacency, applies to larger tree structures and can therefore not be implemented directly in GPSG. Instead, we might explain the ungrammaticality of (3) in terms of a local filter proposed by Horrocks (1987). Figure 5.6 depicts this local filter which prohibits a gap from passing through $\overline{\overline{V}}$ when the latter is the complement of a lexical head noun. This filter will prevent the gap from being discharged in (3), as can be seen from Figure 5.7., thus making (3) ungrammatical. Consequently, as also pointed out by Harris (1991), Slobin (1991), and Wasow (1991), it might be the case that many of the universal principles of UG are mere theoretical artifacts of a particular (Chomskyan) perspective on language. This, in turn, suggests that other linguistic frameworks might lead to significantly different language universals. Nevertheless, what has to be kept in mind concerning the present perspective is that linguistic facts, such as, e.g., the unacceptability of (3) compared with acceptability of (1), may be explained

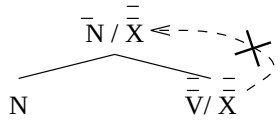


Figure 5.6. A local grammaticality filter preventing gaps (the SLASH feature in GPSG terminology) from passing through $\overline{\overline{V}}$ when the latter is the complement of a lexical head noun (Horrocks, 1987)

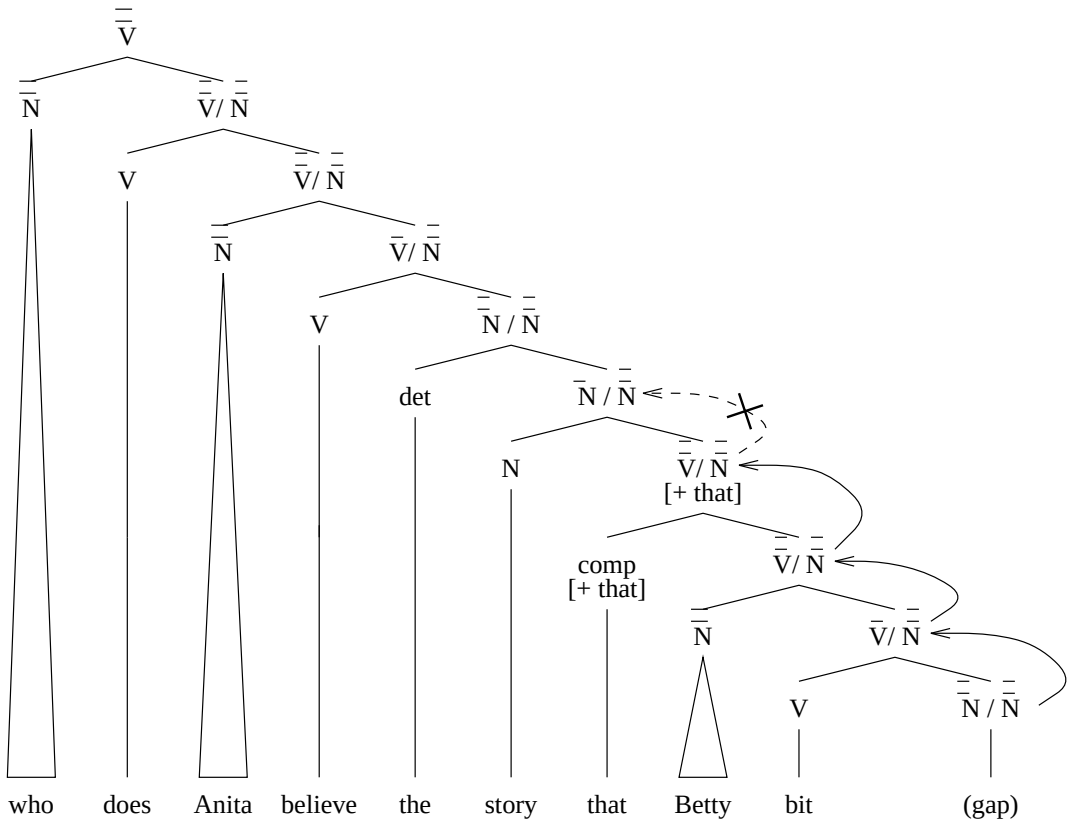


Figure 5.7. A syntactic tree derived via GPSG for the ungrammatical sentence *'Who does Anita believe the story that Betty bit'*. The plain arrows illustrate the upwards percolation of the gap to the point where it is stopped by the ungrammaticality filter of Figure 5.6 (the dashed arrow), preventing the gap from being discharged.

in terms of constraints arising from processing and learning independently of any particular linguistic framework. If one adopts the linguistic perspective of GB, subjacency can be explained as a processing constraint (as we saw earlier). Even though no one yet, to my knowledge, has tried to demonstrate this for GPSG based parsing, local grammaticality filters, such as Figure 5.6, might plausibly follow from a GPSG parser, given the right processing limitations restricting the upwards percolation of gaps.

In closing this section, we might ask whether language is an organ as suggested by Chomsky (1965, 1980, 1986, 1988). I would say 'no'. Rather, we should construe language as an organism with its own evolutionary history (first suggested by Darwin, 1900). This will provide a better understanding of why language looks the way it does today. Nonetheless, it might be objected that Pinker (1994) also invoked Darwin as support for his notion of language as an instinct. Darwin characterized language as a 'wonderful engine' of great importance to human evolution: "A great stride in the development of the intellect will have followed, as soon as the *half-art and half-instinct of language* came into use; for continued use of language will have reacted on the brain and produced an inherited effect; and this again will have reacted on the improvement of language" (1900: p. 634; my emphasis). Elsewhere, he emphasized that "it certainly is not a true instinct, for every language has to be learned" (Darwin, 1900: p. 101). So, as Pinker (1994: p. 20) also seems to acknowledge, Darwin suggested that evolution endowed us with an *instinct to learn language*, rather than a language instinct *per se*. Thus, I submit that it is most fruitful to construe language as a nonobligate symbiant (and not as an organ nor as an instinct). Next, we shall see that this perspective helps us understand how language—as an organism—might have evolved in close relationship with the evolution of its human hosts.

5.2 The Origin of Language

As indicated in chapter 1, the origin of language has been a controversial topic for some time—a topic that still incites much debate (e.g., Bloom, 1994; Corballis, 1992, 1994; Greenfield, 1991; Hurford, 1991; Kiparsky, 1976; Lieberman, 1973, 1976; Piattelli-Palmarini, 1989; Pinker, 1994; Pinker & Bloom, 1990). In this section, I outline a picture concerning the origin and evolution of language. All research on this topic will necessarily be somewhat speculative since language did not leave any fossils to study. Nevertheless, I believe that my sketch receives credibility from being based on theories and data from a variety of fields including historical linguistics, anthropology, implicit learning theory, evolutionary theory (and artificial life simulations thereof), speech perception and production, neuroscience, and connectionist modeling.

Cognitive behavior—human and non-human—can conveniently be divided into two basic kinds of processes: 1) processes dealing with sequential/temporal information in the input[11]; and 2) processes dealing with the categorization and identification of input (or parts of it). In organisms with stereoscopic vision, for example, visual processing is divided into processes dealing with motion detection (that is, sequential changes in the visual array) and processes dealing with object recognition (Bruce & Green, 1985). Motion detection needs to be fast (so that, say, approaching predators can be detected quickly) and therefore only deals with a very reduced part of the visual array. To recognize objects, on the other hand, more detailed information is necessary, leading to slightly slower processing following the heavier information load. Moreover, the two kinds of processes are not completely isolated from each other. They, at least, interact at some level, as when we move our head sideways (thereby varying the angle of the visual input) in order to determine the depth of an object. At other times, we are not able to even detect something (and therefore identify it) before it moves as in the case of a well-camouflaged animal lying still. The distinction also has close analogues in, e.g., the neuropsychological literature (Kolbe & Whishaw, 1990) with its distinction between declarative and procedural types of memory corresponding, respectively, to categorial and sequential processing.

The processes involved in vision are presumably to a large extent hardwired into the brain in order to maximize processing speed at the cost of flexibility in primary visual processing. This trade-off is possible given that the fundamental structure of visual world (i.e., the existence of lines, edges, etc.) does not change dramatically over phylogenetic time, but follow natural laws. In other parts of cognition, flexibility is of the greatest importance. For example, when *Homo erectus* spread from Africa to much of the Old World, it needed to be flexible enough to classify a vast range of new animals in terms of categories, such as PREDATOR, FOOD, etc. Today, we can also find this kind of flexibility in many other parts of cognition, but here I will concentrate on language and speech processing. Liberman & Mattingly (1989), for instance, have demonstrated that the processes underlying the perception of consonants and vowels are distinct from those processes that pinpoint the source of the sound and its auditory qualities. The former processes rely on sequential properties of the input to uncover the appropriate phonological information, whereas the latter categorizes parts of the input stream, in terms of pitch, loudness and timbre, mostly independently of sequential information. Importantly, they stress that the extraction of phonetic information from sequential

[11]Note that input can come either from the environment external to the organism via its sensory apparatus, or from within the organism itself in terms of, for instance, intermediate processing results or feedback from internal organs.

input (i.e., the detection of changes in vocal tract resonances—formants—caused by a speaker's shifting configurations of articulators) has to be learned. So, although all humans are born with the same set of phonetic units, covering the sounds found in all human languages, infants rapidly (within 6 months) acquire a special sensitivity to their native tongue (Kuhl, Williams, Lacerda, Stevens & Lindblom, 1992—more about this later).

More generally, language processing involves the extraction of information occurring in temporal sequences. If learning is involved in language acquisition (a point I will return to, and argue for, at length below), it is most likely that the former will consist in some kind of encoding of sequential structure. Since language acquisition furthermore occurs largely without much conscious effort, it seems reasonable to assume that such learning is *implicit* in the sense of being "an inductive process whereby knowledge of a complex environment is acquired and used largely independently of awareness of either the process of acquisition or the nature of that which has been learned" (Reber, 1992: p. 33). As such, this process may yield abstract information about underlying statistical regularities in the input. Notice, however, that simple co-occurrence may not be sufficient to establish learning (although co-occurrence between stimuli might be recorded in ways which cannot be detected in the behavior of the organism). Instead, it seems necessary that "organisms key on the *covariations* between events and, hence, learn to take advantage of the cuing function that emerges when one event is *contingently associated* with other events. So long as some stimuli in the environment are arranged so that their occurrences cue the occurrences of other stimuli, they will acquire statistical predictive power" (Reber, 1992: p. 45; my emphasis). It is my contention that language learning may be construed in terms of such implicit processing—a point also made by Reber (1992) and Durkin (1989). The simulations presented in the previous two chapters appear to corroborate this view (which the work by Cleeremans, 1993, on connectionist models of sequence processing hints at too).

If the learning and processing of sequential information constitutes some of the most basic elements of cognition—as I have suggested—then we might expect them to have a long phylogenetic past. Indeed, this idea has been advanced recently by Reber (1990, 1992), suggesting that implicit learning processes are evolutionary ancient. Evidence for this suggestion can be found in the fact that these kind of processes are found in organisms from all over the animal kingdom. Sequential learning has been observed not only in other primates (e.g., cf. results reported in Greenfield, 1991, on chimpanzees solving hierarchical tasks involving deliberately sequenced behaviors), but also in rats (demonstrating that they are able to encode the sequential structure following successive trials in maze, Capaldi & Miller, 1988). Another indication of the purported

primary status in cognitive behavior can be found in the robustness of sequence processing in the face of disorders. For example, Abrams & Reber (1988) demonstrated that a group of brain-damaged psychiatric patients performed significantly worse than a group of college students on an explicit learning task (memorizing simple letter-to-number rules), whereas both groups had the same level of performance on an artificial grammar learning task. Finally, the obvious fact that children are able to rapidly and effortlessly acquire a vast variety of (implicit) information (linguistic, cultural, social, and otherwise) during their first years of living despite their shortcomings in explicit learning tasks, such as conscious memorization, suggests that implicit learning is largely age-independent (this is further supported by data listed in Reber, 1992). The bottom-line, following these properties of the implicit learning of sequential information, seems to be that such processes possibly evolved very early in phylogenetic history in order to become so prevalent in human and animal cognition.

From the present viewpoint of language learning we might ask: if this kind of learning is so ubiquitous in nature, why is it the case that no other organism appears to have developed a communication system comparable with human language in its complexity? The answer is related to the larger, more intricate brain of humans compared with all other animals:

> with increasing neurological sophistication, organisms become capable of detecting more and more tenuous covariations. Organisms as primitive as *Aplysia* [a marine mollusk which responds to conditioning] require unambiguous and non-varying pairing of stimuli for learning; humans are capable of the detection of much more subtle statistical covariations involving many stimulus elements and environmental properties. Or, if one prefers other terminology, humans are capable of learning 'rules'. (Reber, 1992: p. 45; my comment)

Now, Reber seems to suggest that the sequential learning mechanisms found in different species are *homologues*; that is, they all date back to a vital adaptation in a very early ancestor of all subsequent organisms capable of sequence processing. However, it is, at least, conceivable that in some species this kind of processing ability arose independently—as *analogues*—simply because these organisms were submitted to essentially the same environmentally pressures, leading to the same solutions (but perhaps implemented in different ways). It does not matter for the purpose of the present argument whether the capacity for sequential learning is subserved by phylogenetically homologue or analogue structures. What is important, however, is that selectional pressures have forced the evolution of sequential learning mechanisms in a multitude of different organisms. This will set the stage—in terms of learning mechanisms—for my account of the origin and evolution of language.

5.2.1 The Birth of Language

The first step towards language in humans might have been, as we shall see, a byproduct of bipedalism[12]. When the hominids split from the apes somewhere between 4 and 8 million years before present, the defining characteristics of the first known hominid, *Australopithecus afarensis* was bipedalism (although it probably often still used its hands in locomotion). Notice that bipedalism not only minimized energy loss during the hottest hours of the day (because less body surface is exposed to the sun when walking upright), but it also freed the hands and arms from participation in locomotion. This hominid later evolved into two distinct lineages known as *robust* and *gracile*. The former lineage eventually became extinct, whereas it is believed that the latter evolved into the *Homo* lineage about 2 million years before present. *Homo habilis* is the earliest known *Homo*, appearing between 2.2 and 1.6 million years ago in Africa. Following this hominid we find the larger *H. erectus*—which spread out of Africa to most of the Old World—and appears to have evolved into archaic *H. sapiens*. Between 200,000 and 150,000 years ago, the anatomically modern form of humans, *H. sapiens sapiens* emerged in Africa. These early humans migrated to the Old World about 100,000 before present, where they gradually replaced the hominids that had migrated earlier. Some 35,000 years before present the last of the early migrants, the Neanderthals, disappeared, leaving *H. sapiens sapiens* behind as the only living hominid.

But when did language evolve? It is likely that the evolution of language started with (or, at least, did not 'take off' until) the emergence of the *Homo* lineage. Endocasts of *H. habilis* seem to indicate the presence of brain structures homologous to purported language areas in modern day human brains, such as Broca's area. This is important since it has recently been suggested that an area located within the left ventral frontal region of the cortex subserves hierarchically organized sequential behavior. Greenfield (1991) argues that "during the first two years of life a common neural substrate (roughly Broca's area) underlies the hierarchical organization of elements in the development of speech as well as the capacity to combine objects manually, including tool use" (p. 531). She presents experimental data from early language learning and from a task involving the hierarchical nesting of cups to support her position. Summing up the results from these experiments, she writes: "from about 9 to 20 months of age, children pass through parallel and quite synchronous stages of hierarchical complexity

[12]The following paragraph briefly outlining the descent of humans from the apes is based largely on Corballis (1992). Notice that these matters are the subject of some debate and should therefore not be regarded as definitive. Still, any changes that the future might bring are not likely to be problematic for my account of language origin, since the latter does not rely directly on these, admittedly, controversial historical data.

in forming spoken words and combining objects" (1991: p. 539).

This account receives further support from a reported study of adult aphasics; that is, individuals with acquired language disorders following lesions in and often closely around Broca's area. A large number of such Broca's aphasics suffer from agrammatism. Their speech lacks the hierarchical organization we associate with syntactic trees, and instead appears to be a collection of single words or simple word combinations. Interestingly, Grossman (1980) found that Broca's aphasics, besides their noticeable agrammatism, also had an additional deficit in reconstructing hierarchical tree structure models from memory. He took this as suggesting that Broca's area subserves not only syntactic speech production, but also functions as a locus for supramodal processing of hierarchically structured behavior. It is, however, possible that the supramodal deficit following lesions to Broca's area might be a result of damage to two distinct sets of cortical pathways, emanating in close proximity from this area of the brain. Indeed, Greenfield (1991) has hypothesized (based on neurobiological and behavioral data) that starting at the age of two a cortical differentiation of Broca's area begins, leading to distinct capacities for linguistic processing and more complex object combination. Whether this hypothesis is true is a matter of some debate (see the comments to Greenfield, 1991, for instance, Jacobs, 1991), but it suffices for the present account that Broca's area appears to be "a multifunction organ adapted to the regulation of sequential activity in several different domains" (Lieberman, 1991: p. 567).

Returning to *H. habilis*, and the first indication of the existence of Broca's area in hominids, we can plausibly endow this hominid with the prerequisites for some kind of language, based on an ability to deal with sequential information[13]. Language might therefore have originated with this hominid as a kind of systematic set of manual gestures (hence the importance of the freeing of the hands following bipedalism cf. Corballis, 1992). However, I contend that this kind of manual language is closer related to presentday sign language than to the gestures we sometimes still use when communicating. The reason for this is that the neurological seat of the production of these early hominid gestures are likely to have been Broca's area in the left hemisphere of the brain (following the discussion above). Moreover, neurological studies of congenitally deaf people have shown that their production of sign language—as in the case of speech in hearing people—can be located to Broca's area, whereas gesturing generally

[13] At this point, it is worth noting that purported homologues to Broca's area have been found in macaque and squirrel monkeys, but that lesioning that area does not significantly impair the monkeys' vocal utterances (cf. Hauser, 1991). Thus, even though these monkeys appear to have homologues to Broca's area they do not seem to use it for vocalization. As Pinker (1994) adds, this could also have been the case for the early hominids, such as *H. habilis* . However, this point only effects the timing of language's first appearance, not the chronological order of its evolution.

is located in the right hemisphere (Neville, 1993). I therefore propose that the function of Broca's area in this early period of hominid evolution was biased towards relatively simple, sequential manual combination (perhaps only slightly more complex than the sequential behavior observed in presentday chimps by, e.g., Matsuzawa, 1991).

As a means of communication, the early hominid manual language had some obvious limitations. For communication to take place at all, the hominids would have had to face each other; for example, making it difficult to communicate during the locomotion of a hunt. Instruction in various manual tasks would also have been difficult. Consequently, it seems plausible to assume that over evolutionary time, besides increasing in complexity, gestural language eventually became augmented with verbal sounds. This was partly subserved by a significant increase in brain size (in fact, a doubling of the brain size took place in the time between *H. habilis* and late specimens of *H. sapiens*, cf. Corballis, 1992). However, the configuration of the vocal apparatus in most of the *Homo* lineage might have constituted an obstacle for the further evolution of language towards its present (predominately) vocal form.

A gradual change of the supralaryngeal vocal tract started some 250,000 years before present—eventually leading to the unique vocal tract of modern humans—with the Broken Hill fossil skulls of *H. sapiens sapiens* dating from 150,000 years ago indicating an intermediate form (Lieberman, 1973). During this period of change, the larynx descended into the pharynx, creating an upside-down L-shaped tract in which the tongue moved backwards in the oral cavity. This permitted modern humans to generate a variety of vocal tract configurations, some of which produced a number of new sounds, such as, vowels like [a], [i], and [u] as well as velar consonants like [g] and [k]. The novel vowels, in particular, have been shown to be extremely valuable in verbal communication since "a normal human speaker can produce the acoustic signals that specify these sounds without being very precise as he maneuvers his tongue" (Lieberman, 1976: p. 668). As a result of the evolution of the supralaryngeal vocal tract, *H. sapiens sapiens* was not only likely to have had a larger phonetic vocabulary than previous hominids, but its vocal utterances also had a much higher degree of acoustic stability, which, in turn, allowed more intelligible speech[14]. Together, these properties of the modern human vocal tract may have paved the way for the transition from a

[14] *H. sapiens neanderthalensis* provides an interesting example of a *Homo* lineage who did not evolve the modern human vocal tract. In fact, their vocal tract is very similar to that of presentday chimpanzees (and of newborn human infants up to an age of about 6 months). Computer simulations reported in Lieberman (1973) suggest that both Neanderthals and chimps would not be able to produce the vowels [a], [i], and [u]. Nevertheless, the former presumably had some kind of vocal language—albeit a language with phonetic deficits relative to that of *H. sapiens sapiens*—perhaps still combined with extensive use of manual gestures.

language system primarily relying on manual gestures to one predominately based on vocal utterances[15].

5.2.2 The Baldwin Effect Revisited

So far, I have argued that language may originate early in the *Homo* lineage, arguably subserved by neurological structures (in the vicinity of Broca's area) adapted to the learning and processing of sequential information. As argued in the start of the present section, these processes are likely to be evolutionary ancient, but have evolved to deal with highly complex hierarchically organized input (such as, natural languages) through human evolution. At this point, it might be possible to (more or less) accept the above evolutionary scenario concerning the origin of language, but then argue that further evolution has worked to make language innate as hypothesized by the theory of UG. For example, although Pinker & Bloom (1990) acknowledge that some kind of learning mechanism not specific to language may have been underlying the origin of language, the (earlier mentioned) Baldwin effect (Baldwin, 1896) would gradually have caused the language acquisition process to become entirely language specific.

To support this contention, Pinker & Bloom rely on a simulation reported by Hinton & Nowlan (1987) and discussed in Maynard-Smith (1987). Recall that the idea of the Baldwin effect is to allow phenotypical learning to influence the genotype (in a nonLamarkian way) and thereby accelerate evolution. In their simulation of the Baldwin effect, Hinton & Nowlan (1987) investigate the evolution of a very simple organism consisting of a neural net with 20 possible connections. A 20 bit vector designates the organism's genotype in which each bit corresponds to a gene containing one of three different alleles: 1, 0, or ? (the first, specifying the presence of a connection; the next, the absence of one; and the last, an open/close switch to be set via learning). For the organism to increase its adaptive fitness, all connections must be set correctly. Hence, it is no more advantageous to have 19 correctly set connections than just 1. The genotypes of the first 1,000 organisms, which made up the initial population, had on average 10 alleles specifying learnable switches (?) and 5 of each of the two remaining alleles for, respectively, the presence (1) or absence (0) of connections. Each organism is allowed 1,000 learning trials (each a random setting of the switches) in its lifetime. When an organism hits on a switch combination, which, together with the right settings of the genetically specified connections, is identical with the unique adaptive configuration,

[15]Incidentally, Corballis (1992) has suggested that this 'second freeing of the hands' lead to the explosion of cultural artifacts about 35,000 years ago. "That is, manufacture and language would no longer be competing for the same medium, and both could then develop without mutual interference" (p. 214).

learning stops. Notice that only if the *all* genetically specified genes have the correct alleles, is it possible for an organism to increase its fitness through a correct setting of the switches. Procreation involves 1,000 (crossover) matings between two parents, each selected according to their Darwinian fitness (which is determined as a probability proportional to the number of learning trials (n) remaining after the right configuration has been attained: $1 + 19n/1000$). In this way, the sooner an organism reaches the adaptive configuration, the more likely it is to be chosen as a parent (i.e., the better reproductive fitness).

During the first 10 generations, selection did not seem to have much impact, but this changed within the next 10 generations. After 20 generations, the incorrect alleles have all but disappeared from the genotype, leaving all genetically specified alleles as being correct. Hence, it can therefore be said that learning has guided the evolution of this organism (just as the Baldwin effect suggests). Unfortunately, Hinton & Nowlan make a too strong remark regarding the amount of genetic hardwiring obtained in the simulation: "there is very little selective pressure in favor of genetically specifying the last *few* potential connections, because a few learning trials is almost always sufficient to learn the correct settings of just a *few* switches" (1987: p. 497; my emphasis). In fact, the last 'few' connections that needs to be learned amount to about 45% of the total number of alleles, whereas the remaining 55% of these are specified genetically (an absolute difference of 2 alleles). It is clear that the approximately 25% incorrect alleles that an organism starts out with, disappear; and that the number of correct alleles, on the other hand, more than double from about 25% to 55%. It seems to be the case that when an organism has a certain learning capacity, the latter might allow the former to change maladapted parts of the genotype towards a better evolutionary fit, while still maintaining much of the original learning ability. This point is corroborated by recent simulations by French & Messinger (1994), showing that if the phenotypical plasticity of a given trait is high (i.e., most individuals are able to acquire it), then it is less likely to become innate (via the Baldwin effect).

Now, what the simulations by Hinton & Nowlan (1987) and French & Messinger (1994) suggest is that when an organism brings powerful learning mechanisms to bear on the acquisition of a particular task, the Baldwin effect might, perhaps, lead to somewhat more biased learning mechanisms, but not to the high degree of task specificity associated with innately specified acquisition processes. From its origin, as we have seen earlier, language has been subserved by sequential learning mechanisms of considerable power, suggesting that the Baldwin effect might be an unlikely explanation of how the purported UG became innate. An additional argument in support of this suggestion can be found by looking closer at one of the assumptions underlying the Baldwin effect:

"In a fixed environment, when *the best thing to learn remains constant*, this can lead to the genetic determination of a character that, in earlier generations, had to be acquired afresh each generation (Maynard-Smith, 1987: p. 761; my emphasis). But given the present perspective on language as a nonobligate symbiant, whose rate of evolutionary change is much higher than that of its hominid hosts, it is clear that 'best thing to learn' does *not* remain constant. That is, the hominid learning mechanisms involved in language learning are 'chasing' a continuously moving target. Of course, it is possible to object that some properties of language will be stable across linguistic change, and that these properties could eventually become innate via the Baldwin effect. However, if the present account of language evolution is correct, then whatever stable (or universal) properties we might find, they are going to be—first and foremost—byproducts of the learning and processing of language. As such, language has adapted to the idiosyncrasies of the hominid learning and processing mechanisms subserving language. The subsequent universal properties have therefore been a 'part' of, or rather, artifacts of the hominid genotype from the very beginning of language evolution. The upshot seems to be that the Baldwin effect cannot do the job of producing the massive innate language endowment assumed by UG (pace Pinker & Bloom, 1990), leaving much room for genuine language learning in modern humans (a case for which I present additional developmental arguments in section 5.3).[16]

5.2.3 Linguistic Change

Having provided an evolutionary account of the development of the ecological niche within which language has evolved—its hominid hosts—we can now turn to the evolution of language itself as evidenced through linguistic change. In unison with Corballis (1992), I contend that early vocal language started out in a very primitive form, perhaps involving only a single vowel. The changes in the supralaryngeal vocal tract following the emergence of *H. sapiens sapiens* some 150,000-200,000 years before present, allowed the development of more sophisticated forms, with vowel differentiation and better articulation of consonants. I propose that some very rudimentary syntax might have been the next evolutionary step, permitting the description of actions involving a subject and perhaps an object. This kind of early language presumably did not involve a morphological system; rather, single and plural instances of a particular thing

[16]This leaves Pinker & Bloom (1990) and Pinker (1994) with a problem in their evolutionary explanation of UG. Either they have to bite the bullet and admit that an innate, language specific system, such as UG, could not have evolved through natural selection augmented by the Baldwin effect (and, subsequently, that much learning still takes place in language acquisition); or, they would have to come up with a different evolutionary story explaining how language eventually became innate.

(or event) were simply referred to by different words (perhaps with some minor, but unsystematic, phonological similarities).

Up until about 35,000 years before present, I hypothesize that the evolution of language was quite slow, perhaps only involving the addition of relatively few vocabulary items, but following the explosion of cultural artifacts from this point in human evolutionary history, we would have seen a dramatic growth in the number of words (referring to the multitude of new artifacts and other aspects of the rapidly evolving culture), and perhaps, a more complex syntax allowing for the expression of more complex situations and events. The key to understanding the subsequent evolution of language, I submit, is vocabulary growth within a system of limited learning and processing capacity. To accommodate the spurt in vocabulary growth, language had to develop morphological systems in order not to exceed the learning and processing capacities of its human hosts[17]. At first, these early morphological systems would have been quite complex, but later developments would have simplified them gradually, so as to incorporate an ever growing vocabulary within the limits of human memory. It is likely that syntax would have become more complex as the morphology became simpler. Wang (1976), in a discussion of a study of relative clause formation in English over the past 1,000 years, writes that "the simplification of morphology ...although it makes the language easier to learn, has the adverse effect of leaving too many sentences unmarked at their clause boundaries. In order to reduce ambiguities for the listener, relative clauses become better and better marked" (p. 65). This explains the appearance of generally stronger restrictions on relative clause markers and the disappearance of inflections first in nouns and, subsequently, in verbs[18].

The proposed effects of human learning and processing constraints on linguistic change is further substantiated by recent simulations by Hare & Elman (1994), investigating the changes in English verb inflection over the last 1,100 years. The morphological system of Old English (ca. 870) was quite complex involving at least 10 different classes of verb inflection (with a minimum of six of these being 'strong'). The simulations involved several 'generations' of neural networks, each of which received as input the output generated by a trained net from the previous generation. The first net was

[17] This development is somewhat similar to the shift from rote-learning to system building evidenced in the acquisition of vocabulary items in early childhood and, importantly, also observed in a connectionist model of this acquisition process (Plunkett & Marchman, 1993).

[18] Here it should be noted that—as pointed out by Kiparsky, 1976—there appears to be cases where a language has changed towards a more complex morphology (presumably with no significant change in syntax). Such changes often occurs in relatively isolated cultures where the influx of new words is low and overall vocabulary growth may be rather slow. These cases therefore do not pose a problem for my account, since it focuses on vocabulary growth, but also can allow for the possibility that lack of vocabulary growth can lead to an increase in morphological complexity over time.

trained on data representative of the verb classes from Old English. However, training was stopped before learning had reached optimal performance. This is meant to reflect a difficulty in learning the inflected forms to perfection. In the present framework, we can construe this difficulty as a consequence of a vocabulary whose requirements, in terms of memory processing, is close to the limits of its human hosts. The imperfect output of the first net is used as input for a second generation net, for which training is also halted before learning reaches asymptote. Output from the second net is then given as input to a third net, and so on, until a total of seven generations has been trained. This training regime leads to a gradual change in the morphological system being learned; a change following imperfect learning of the inflection of verbs that either have a low frequency of occurrence in the training set, or have little internal phonological consistency (i.e., verbs placed far in phonological space from any prototype). Their results show that "verbs which either belong to small classes, lack consistent defining characteristics, or are low in frequency should change most rapidly; change in other verbs will depend on the precise extent to which they possess the characteristics which make them resistant to assimilation" (Hare & Elman, 1994: p. 31). The change taking place in the verb inflection system being transferred between generations closely resembles, in considerable detail, the historical change in English verb inflection leading from the highly complex past tense system of Old English to the modern English incorporating one dominating class of 'regular' inflected verbs and a small set of 'irregular' verbs.

In this section, I have provided an account of the origin and evolution of language, pointing to sequential learning processes as possible mechanisms for language learning and their possible underlying neurological substrate; presenting arguments against language becoming innate via the Baldwin effect; and, briefly outlining a scenario describing the emergence and subsequent linguistic evolution of vocal language (from manual language). Notice that both the exaptationist (e.g., Chomsky, 1972, 1982, 1988, 1993; Piattelli-Palmarini, 1989) and the adaptationist (Pinker, 1994; Pinker & Bloom, 1990) UG positions (mentioned in the previous section) seem to be hard pressed when it comes to explanations of language change, whereas the present account of language evolution, as we have seen, lends itself easily to such explanations (perhaps, but not necessarily, couched within a connectionists framework). For example, Pinker (1991) has put forward a dual-route model of modern English noun/verb morphology, suggesting that irregular forms are stored in a neural net-like fashion, whereas regular inflection is produced via a rule component adding the prefix '-ed'. However, this kind of model is faced with the problem of providing an account "for the qualitative shift from a system with many verb classes of roughly equal status to one with a single

rule and scattered exceptions, in order to explain how the current regular/irregular system developed" (Hare & Elman, 1994: p. 34–35). The model furthermore suffers from an ontogenetic version of this problem which has to do with the explanation of how the rule component develops in early language acquisition. On the other hand, Plunkett & Marchman (1993) have presented a connectionist account in which a single mechanism explains the qualitative shift from rote-learning to system building in the early acquisition of vocabulary items. On this developmental note, I will now turn to the ontogenetic development of language as reflected by the above account of its phylogenetic history.

5.3 Language Learning, Maturation, and Innateness

As should be clear, the account of the origin and evolution of language presented in the previous sections favors a learning based view of language acquisition—a viewpoint which is significantly different from present UG inspired perspectives. However, it would seem that such an account would be impossible given the apparent *poverty of the stimulus* (for instance, cf. Chomsky, 1986, 1993; Crain, 1991; Piattelli-Palmarini, 1989, 1994; Pinker, 1994; Pinker & Bloom, 1990). The primary linguistic input, which is available to a child acquiring language, appear to be so noisy and provide such a poor generalization basis that learning seems almost inconceivable. This leads to the following language learning paradox: On the one hand, learning a language involves deriving a complex model of language structure, given noisy and partial input, apparently without the benefit of usable feedback from others, making this an extraordinarily hard task. Children, on the other hand, are capable of acquiring language rapidly and routinely at a time when their remaining cognitive abilities are quite limited. A possible solution to this paradox, is to suggest that "in certain fundamental respects we do not really learn language; rather, language grows in mind" where 'learning' is understood as being processes of "association, induction, conditioning, hypothesis-formation and confirmation, abstraction and generalization, and so on" (Chomsky, 1980: p. 134–135). In this section, I will present an processing based account of language learning in which a reappraisal of the poverty of the stimulus argument will pave the way for a solution to the language learning paradox.

Based on evidence from studies of both first and second language learners, Newport (1990) has proposed a *"Less is More"* hypothesis which suggests "paradoxically, that the more limited abilities of children may provide an advantage for tasks (like language learning) which involve componential analysis" (p. 24). Maturationally imposed limitations in perception and memory forces children to focus on certain parts

of language, depending on their stage of development. Interestingly, it turns out that these limitations make the learning task easier because they help the children acquire the building blocks necessary for further language learning. In contrast, the superior processing abilities of adults prevent them from picking up the building blocks directly; rather, they have to be found using complex computations, making language learning more difficult (hence, the notion of a crucial period in language learning, a point we shall return to). This means that "because of age differences in perceptual and memorial [sic] abilities, young children and adults exposed to similar linguistic environments may nevertheless have very different internal data bases on which to perform linguistic analysis" (Newport, 1990: p. 26).

In relation to morphology, Newport discusses whether a learner necessarily needs *a priori* knowledge akin to UG in order to segment language into the right units corresponding to morphemes. She finds that this is not the case; rather, segmentation may, indeed, be possible "even without advance knowledge of the morphology, if the units of perceptual segmentation are (at least sometimes) the morphemes which natural language have developed" (1990: p. 25). More recently, Goldowsky & Newport (1993) have corroborated this point via a computer simulation of the acquisition of a simple artificial system of morphology. When comparing a system receiving unrestricted input with a system equipped with a restrictive input filter—whose constraints on the input was gradually loosened over time (to simulate maturation)—they found that the former fails to learn an efficient representation of the data, whereas the latter is able to acquire an optimal solution to the morphological mapping task. These simulation results therefore lend further support to Newport's (1990) less-is-more hypothesis.

The less-is-more hypothesis has a natural interpretation within the present framework. As mentioned earlier, language—being a nonobligate symbiant—confers selective advantage onto its hominid hosts through its function as a means of communication. To reap the full advantage of language, it is important for the hominids to acquire language as early as possible in their life time (so as to have the increased fitness for the longest time possible). This means that for language to be the most useful to its hosts it must adapt itself in such a way that it is learnable by children despite their limited memory and perception capacities. Importantly, this imposes strong constraints on what language can look like, since languages that are not learnable by children will disappear. That is, languages have evolved to be learnable primarily by children which explains why acquiring a language as an adult can be quite difficult. The critical period can therefore be viewed as a spandrel with no particular adaptive properties, emerging as a by-product of language adapting to the learning and processing capacities of infant hominids.

5.3.1 The Critical Period

Hurford (1991) also finds the critical period to be a spandrel; albeit, he arrives at this conclusion coming from a significantly different direction. In a computer simulation of some complexity, Hurford demonstrated that natural selection in a period of 1,000 generations would push a population of individuals from having no knowledge of language at all to one with an ability for the acquisition language early in life. Notice that in this framework, the critical period implies the ability to acquire language ceases to be an advantage at later life stages. Hurford suggests that "thinking at the end of the critical period as 'switching off', like the deliberate switching off of a light, is less appropriate than thinking of it as a point where the 'energy' in the system, the selection pressure in favor of positive alleles, is dissipated, and the 'light' goes out for lack of pressure to keep it 'on' " (1991: p. 193). However, it is worth noting that each gene in the simulation can have either an allele which inhibits language acquisition, or one which facilitates it. Regarding the former, Hurford argues that the inhibitive genes serves as a kind of trade-off cost, simulating that adaptation for a particular trait might inhibit other traits, such as, for example, when a sea turtle's adaptation for swimming inhibits its locomotion on land. It is clear that some trade-offs have been made in the evolution of language. This appears to be the case with the earlier mentioned change of the supralaryngeal vocal tract which, in its modern adult form, significantly inhibits respiration, decreases our sense of smell, and increases the possibility of choking on food (Lieberman, 1973, 1991). Nevertheless, I question Hurford's particular implementation of this evolutionary principle insofar as this is arguably the dominating factor leading to the apparent evolutionary disadvantage of language acquisition in later life stages. It is not clear that the language acquisition 'light would go out' if trade-off was implemented in a different (perhaps more realistic) fashion.

More recently, Pinker (1994) has picked up on Hurford's simulation results, suggesting to conceive maturational changes in terms of "a machine shop in a thrifty theater company to which props and sets and materials periodically returns to be dismantled and reassembled for the next production" (p. 294). Consequently, this picture suggests that once language has been acquired the language acquisition machinery should be sent back to the machine shop for recycling (since the maintenance of neural tissue is very resource demanding). Although as an evolutionary point this is perhaps conceivable, the theater machine shop metaphor does not make much sense in terms of brain maturation. What would the recycling of large areas of neural tissue amount to in the development of the brain? Granted the brain has a high degree of plasticity (see, e.g.,

Ramachandran, 1993; Recanzone & Merzenich, 1993), but it is unlikely that it is capable of the kind of restructuring required by the Pinker's scenario. Furthermore, the metaphor relies, I think mistakenly, on the existence of a language acquisition device functionally separate from whatever machinery is involved in the processing of language. Such functional separation of the learning and processing of language might be a historical left-over from Chomsky's (1965) idealization of instantaneous acquisition of language, suggesting that the adult language competence can be studied independently of how it it is acquired. In any event, if this separation is valid, we would expect to be able to obtain evidence indicating some kind of dissociation between the ability to acquire and process language; and, to my knowledge, no such data have been reported.

Connectionist models, on the other hand, do (as pointed out in chapter 2) not require a separation of learning and processing. Indeed, they seem to suggest that both must be subserved by the same underlying mechanisms. Now, it is clear that Newport's (1990; Goldowsky & Newport, 1993) research only addresses the acquisition of morphology, but it is clear that on my present account the same kind of explanations would also apply to the acquisition of syntax. Indeed, my view implies that the most—if not all—of the purported universal syntactic principles of UG may not need to be postulated as chunks of innate domain-specific knowledge, but are instead proposed to be mere artifacts of a learning mechanism undergoing maturational development. Such maturational changes are, of course, still innately specified, but, importantly, in a predominately domain-general fashion. Evidence that maturational limitations on memory might facilitate the acquisition of syntax can be found in the simulations presented in chapter 4, extending earlier work by Elman (1991b, 1993). Recall that these simulations showed that when a network was undergoing (simulated) maturational changes (in the form of a gradually loosening constraint on memory, as originally proposed by Elman), it was possible for it to learn to respond appropriately to sentences derived from a small phrase structure grammar of considerable linguistic complexity. However, without this maturational constraint the nets would fail to learn the task (displaying a behavior which might be somewhat comparable with that of late learners). These simulation results support the idea that maturational constraints (of some sorts) on a learning mechanism may allow it to acquire relatively complex linguistic structure without presupposing the existence of innate language specific knowledge[19]. I find that

[19]. Still, it might be objected that these connectionist simulations only deal with small artificially generated corpora and will therefore not be able to scale up to noisy real world data. This might be true, but recent research in statistical language learning suggests otherwise. The earlier mentioned results obtained by Finch & Chater (1993) demonstrated that simple statistics—similar to what the above networks are sensitive to—can filter out noise and induce lexical categories and constituent phrases from a 40 million word corpus extracted from INTERNET newsgroups.

such a maturationally constrained learning mechanism may provide a partial solution to the language learning paradox, but the question still remains of how to overcome the apparent poverty of stimulus.

5.3.2 The Poverty of the Stimulus Reconsidered

The roots of the poverty of stimulus argument goes back about thirty years to Chomsky (1965), but has been elaborated considerably since (e.g., amongst others, by Chomsky, 1986). This argument against language learning, in fact, consists of a several inter-related sub-arguments:

- *Arbitrary universals:* A number of the universal linguistic principles that children appear to acquire are essentially arbitrary and have no direct reflection in the input. For example, a child knows that the earlier mentioned sentence *'Who does Anita believe the story that Betty bit'* is ungrammatical without ever having received information that this is so.

- *Noisy input:* The linguistic information available to a child is considerably noisy, consisting of both grammatical and ungrammatical sentences, but with no additional information about which is which.

- *Infinite generalization:* Children receiving significantly different sets of input are nontheless capable of converging on the same underlying grammatical regularities (if you wish, grammar), serving as the finite basis from which an infinite number of utterances can be produced.

- *Early emergence:* Children are capable of applying a number of complex linguistic principles so early in their language development that it seems unlikely that sufficient time have elapsed for learning to be feasible.

- *Inadequacy of learning methods:* The (empiricist) general-domain learning methods available in the explanation of language learning do not appear to be adequate for the purpose of language acquisition.

Together, these sub-arguments militate—decisively, according to many researchers (e.g., Chomsky, 1986, 1993; Crain, 1991; Piattelli-Palmarini, 1989; 1994; Pinker, 1994)—against learning based approaches to language acquisition, favoring a UG perspective comprising a substantial innate database of linguistic principles and only little, language specific learning. Nevertheless, as will become evident below, each of these sub-arguments become less convincing vis-a-vis the present account of language evolution and development.

The purported *arbitrariness* of universal linguistic principles is an intrinsic part of UG approaches to language (e.g., Crain, 1991; Piattelli-Palmarini, 1989; Pinker & Bloom, 1990). However, if we adopt the perspective on language as a nonobligate symbiant (as suggested in section 5.1) and the subsequent evolutionary picture (outlined in the previous section), then it is not necessary for the child to learn the universal principles. Rather, they come for 'free' given that they are natural side-effects of sequential processing and learning in a system with certain memorial and perceptual limitations (recall the earlier mentioned processing based explanation of the paradigm case of an arbitrary linguistic principle: subjacency). Notice that on this account the universal principles are not language specific chunks of knowledge which have been stored innately (via the Baldwin effect), but by-products of the processing and learning of language by mechanisms evolved for the more general task of dealing with sequential structure in the environment. In other words, the apparent fact that the universal linguistic principles appear to be essentially arbitrary, and, furthermore, cannot reliably be induced from the primary linguistic input, does *not* warrant a dismissal of learning based approaches to language acquisition, if these universals, as I have suggested, can be shown to be mere by-products of learning and processing in a system with certain resource limitations.

It is often noted that the primary linguistic data available to a child is inconsistent and full of *noise*. This is clearly a problem for many classical models of language learning, such as, for example, the parser proposed by Berwick & Weinberg, 1984. In this model, language learning proceeds by attempting to parse input sentences; when parsing fails a new rule is added to the grammar so that parsing can continue. It is not hard to imagine the catastrophical consequences following the presentation of inconsistent input (such as, the ungrammatical sentence *'Who does Anita believe the story that Betty bit'*). The model would add rules which would lead to ungrammatical sentences. Interestingly, connectionists models do not typically suffer from this problem. Indeed, as pointed out by Seidenberg (1994) "it is demonstrably true that networks can learn to solve problems in the face of inconsistent training ...analysis of these learning systems suggests that at least some inconsistencies in feedback or network behavior might actually *facilitate* finding the solution to a problem" (p. 392). Even though it is presently uncertain whether such models can solve the problem of language acquisition under conditions corresponding those facing a child, it is clear that the existence of noise in the primary linguistic input cannot be used as an *a priori* argument against connectionist models of language learning.

The problem of *infinite generalization* given finite input has received considerable attention in the language learning literature (e.g., Elman, 1993; Gold, 1967; MacWhinney, 1992; Pinker, 1984, 1989), and is related to the previous problem of noisy input data. In a now classic paper, Gold (1967) proved that not even regular languages can be learned in finite time from a finite set of positive examples (i.e., grammatical sentences). This proof combined with the lack of observed negative input found in the primary linguistic data (that is, ungrammatical sentences do not come labeled as such, neither do adults reliably correct a child's ungrammatical utterances) leads to a predicament regarding human language learning. Gold therefore suggested that his finding must lead to at least one of the following three suppositions. Firstly, it is suggested that the learning mechanism is equipped with information allowing it to constrain the search space dramatically. In other words, innate knowledge will impose strong restrictions on exactly what kind of grammars generate *the* proper projections from the input to (only) human languages. This is the approach which goes under the name of UG. Secondly, Gold proposes that children might receive negative input that we are simply not aware of. This would allow the correct projection to only human languages. However, see Pinker (1989) for an extensive discussion and subsequent dismissal of such a proposal (though the prediction task, as applied in the simulations reported in chapter 3 and 4, might be construed as a kind of *weak* negative feedback). Thirdly, it could be the case that there are *a priori* restrictions on the way the training sequences are put together. For instance, the statistical distribution of words and sentence structures in particular language could convey information about which sentences are acceptable and which are not (as suggested by, for instance, Finch & Chater, 1993). Regarding such an approach, Gold notes that distributional models are not suitable for this purpose because they lack sensitivity to the order of the training sequences.

So, it seems that *prima facie* UG is the only way to get a language acquisition off the ground—even though learning has to take second place to innate domain-specific knowledge. Nevertheless, given our earlier discussions it should be clear that this conclusion is far from inevitable. The way out of Gold's predicament without buying into UG can best be fleshed out by taking a closer look at the two basic assumptions which the proof is based on:

> Given *any language of the class* and given *any allowable training sequence* for this language, the language will be identified in the limit [i.e, it is learnable in finite time from a finite set of examples]." (Gold, 1967: p. 449; my emphasis and comment).

First of all, Gold is considering *all* possible permutations from a finite alphabet (of words) into possible languages constrained by a certain language formalism (e.g.,

context-free or finite-state formalisms). Thus, he stresses that "identifiability (learnability) is a property of classes of languages, not of individual languages." (Gold, 1967: p.450). This imposes a rather stringent restriction on candidate learning mechanisms, since they would have to able to learn *the whole class* of languages that can be derived from the combination of an initial vocabulary and a given language formalism. Considering the above discussion of language as a nonobligate symbiant, this seems like an unnecessarily strong requirement to impose on a candidate for the human language learning mechanism. In particular, the set of human languages is much smaller than the class of possible languages that can be derived given a certain language formalism. So, all we need to require from a candidate learning mechanism is that it can learn all (and only) human languages, not the whole class of possible languages derivable given a certain language formalism.

Secondly, Gold's proof presupposes that the finite set of examples from which the grammatical knowledge is to be induced can be composed in an arbitrary way. However, if the learning mechanism is not fixed but is undergoing significant changes in terms of what kind of data it will be sensitive to (as discussed above), then we have a completely different scenario. Specifically, even though the order of the *environmentally presented* input that a learning mechanism is exposed to might be arbitrary, the composition of the *effective* training set is not. That is, maturational constraints on the learning mechanism will essentially reconfigure the input in such a way that the training sequence always will end up having the same effective configuration (and this is, in effect, comparable with Gold's third suggestion). Importantly, this is done without imposing any restrictions on the publically available language, i.e., the language that the child is exposed to. This point is further corroborated by simulations presented in chapter 4. There it was found that a network trained while undergoing maturational changes was capable of dealing with sentences derived from a *context-free* phrase structure grammar; and was furthermore able to produce (strong) generalizations.

Having thus 'disarmed' the apparent predicament following Gold's learnability proof, we now turn to the fourth sub-argument: the *early emergence* of linguistic competence. Crain (1991) presents results from a number of psycholinguistic experiments involving children aged between 2 and 5 years, suggesting that they are capable of obeying various complex linguistic constraints at early stages in their language acquisition. He adds that "there has been a steady increase in recent years in the number of empirical demonstrations of children's mastery of syntactic facts for which they have little, if any, corresponding experience" (p. 611). This is then taken as evidence for a scenario in which children are guided through language acquisition predominately by innate, language specific knowledge largely encoded as universal constraints. Nonetheless, the

present learning and processing based account of language acquisition is fully compatible with the same data. In particular, this approach would predict that some universals would be observable quite early in life because they emerge as side-effects of learning and processing; but it also promises to provide a chronological explanation of the differences in the time of onset between the various universals. A closer understanding of how maturational changes affect learning and processing capabilities is hypothesized to provide the basis for such an explanation.

Piattelli-Palmarini (1989) has in a similar fashion contended the existence of an innate UG: "The newborn (and even prematurely born) infant displays a highly sophisticated array of specific linguistic filters allowing for instant discrimination between linguistic and non-linguistic sounds" (p. 27). A closer look at the evidence concerning infant auditory perception, however, suggests that such linguistic filters are learned, not innate. Cross-language experiments investigating infant auditory perception have demonstrated that 6-month-old infants in the United States and Sweden display *'perceptual magnet effects'* specific to their native language: "Linguistic experience shrinks the perceptual distance around a native-language prototype, in relation to a nonprototype, causing the prototype to perceptually assimilate similar sounds" (Kuhl *et al.*, 1992: p. 608). This magnet effect facilitates speech processing by permitting less salient instances of a phonetic category (characteristic of one's native language) to be perceived as a clear instance of that particular category, thus making speech perception more robust (i.e., less prone to phonetic misclassification). Earlier experiments have shown that nonhuman animals (e.g., Rhesus monkeys) do not seem to acquire perceptual magnets (Kuhl, 1991). Nevertheless, both nonhuman animals (e.g., chinchillas and macaques) and human infants have been found to start out with the *same* innate ability to make a number of phonetic differentiations in auditory perception (Kuhl, 1987). These differences appear to correspond exactly to the collection of phonetic units needed to perceive all the world's different languages. Given the present perspective on language evolution, this is not surprising: language must necessarily keep itself within the boundaries of its ecological niche, part of which is made up by the basic auditory perceptional capacities of hominids (and other animals). Thus, "infants' ability to hear the differences between phonetic units is innate and attributable to general auditory processing mechanisms. Perceptual boundaries are not argued to be due to special processing mechanisms that evolved for language in human beings" (Kuhl, 1993: p. 48). That learning is fundamental in speech perception is further supported by experiments showing that 4-day-old infants are able to distinguish utterances from their own native language from those of other languages—even when the speech samples were passed through a (400 Hz) low-pass filter (Mehler, Jusczyk, Lambertz,

Halsted, Bertoncini & Amiel-Tison, 1988). Pace Piattelli-Palmarini (1989: p. 28; note 19), this indicates that learning already starts in the womb (despite distortions brought about by the amniotic fluid), allowing the infant to become sensitive to the prosody of his or her native language.

Finally, we can turn to the last sub-argument concerning the general *inadequacy of (empiricist) learning*. A version of this argument has been presented by Ramsey & Stich (1991) as the 'competent scientist gambit'. The idea of this thought experiment is to replace the language learning mechanism(s) with the best empiricist scientist around. Her task is to come up with the specific grammar underlying a collection of utterances, making up the primary linguistic input from a language unknown to her. She is allowed to use any empiricist technique available, but she is not permitted to learn the language herself since that would allow her to use (her own) grammaticality judgements to determine the grammar (information that is not available to a child acquiring his or her first language). It should be apparent that specified in this way, the task is certainly nontrivial, if not impossible. Given the data available there seems to be no guarantee that she will end up with the right grammar (since there is no way of proving that the final grammar will not overgeneralize). Ramsey & Stich show in a number of refinements of this thought experiment that it appears to be necessary to equip the scientist with information akin to an innate UG before we have a reliable procedure for grammar 'learning'. Arguments of this sort are therefore taken to show that empiricist language learning is impossible even in principle.

I think this argument is misguided for, at least, three reasons. First, there seems to be no reason to expect that whatever structure language might take in our heads (so-to-speak) should be readily available to us by means of introspection. That is, our language acquisition machinery cannot be likened to a homunculus (even if it has the qualities of a competent scientist). Instead, if the present account is right, it is going to be a statistically based learning device[20]. This leads to a second caveat concerning the kind of learning devices the scientist may employ. For example, as argued above, a maturationally constrained learning process might prevent overgeneralization by 'reordering' its effective input. Of course, one could object that Chomsky (1956) demonstrated that statistical approaches to language acquisition are inadequate. Here it is, as Elman (1993) stresses, "important to distinguish between the use of statistics as the *driving force* for learning, and statistics as the *outcome* of learning"(p. 87). Chomsky's proofs was only concerned the latter form of learning, whereas connectionist

[20]Indeed, the experimental results presented in Juliano & Tanenhaus (1993) point in this direction (see also Seidenberg, 1994, for a similar view).

models typically rely the former. The final lacuna in the competent scientist gambit is the assumption that empiricist oriented learning cannot involve significant, but nevertheless non-language specific, innate constraints (or biases) on the learning process. The present account of language learning and evolution challenges that assumption without becoming yet another version of UG. The bottom-line is that the competent scientist gambit intuition pump appear to run out of steam when faced with a statistically oriented language learning account constrained by processing and maturational considerations.

5.3.3 The Objection from Creolization

In the remaining part of this chapter, I will briefly address a couple of possible objections that *prima facie* appear to be rather damaging to a learning based account. The first of these is the *creolization* of pidgin languages (and other languages with a similarly inconsistent structure). Originally, creolization was the name of a process in which children made a pidgin language their own native tongue, creating a creole. A pidgin is a language which lacks any significant grammatical structure and in which communicative understanding therefore must rely mostly on constraints provided by lexical semantics, and perhaps intonation. A creole, on the other hand, displays a clear—albeit, from the standpoint of established languages, not perfect—grammatical structure. Bickerton (1984) reports a case of creolization in a single generation. At the end of the last century, workers from a variety of countries were brought together on Hawaiian sugar plantations. The need for communication soon led to the development of a pidgin. Many of the plantation workers' children spent most of their time in separation from their parents, overseen by a worker speaking pidgin. These children ended up inventing Hawaiian Creole. Another example of (what may be) creolization was reported by Singleton & Newport (1993) in a study of a 7 year old congenitally deaf boy. Simon was born to deaf parents who started to learn American Sign Language (ASL) at about 15 years of age. Simon's only source of input was the 'approximated ASL' of his late-learner parents. In tests comparing Simon's performance on ASL morphology with that of his parents, it was clear that his performance surpassed theirs. Moreover, Simon's performance was shown to be equal to that of children whose parents were native speakers of ASL, despite receiving a more inconsistent set of linguistic input. Simon was thus able to become a native speaker of ASL, something his parents never achieved.

Pinker (1994) and Bickerton (1984) take such creolization as strong evidence for an innate UG, permitting (or rather, forcing) children to impose structure on inconsistent

input, as in Simon's case, or on almost structure-less input, in the case of the children on the Hawiian plantations. It is clear that the kind of 'classical' learning devices that are presupposed in the competent scientist gambit will not be able to do the trick. But does creolization therefore make language learning impossible? Some reason for optimism can be found in Singleton & Newport's (1993) suggestion that "many kinds of devices, not necessarily restricted to language learning, will work like this: sharpening consistent mappings, and ignoring or losing inconsistent ones" (p. 49). One could further hypothesize that if language acquisition is subserved by maturationally constrained learning mechanisms, acquiring sequentially organized hierarchical structure, then such a process might itself impose structure where none is in the input, or where the structure is plagued by inconsistency. Although this account is, admittedly, somewhat speculative, it does, at least, rule out the possibility of any in principle arguments based on creolization against language learning (while pointing towards a possible simulation experiment which will be described in the section on future directions in the final chapter).

5.3.4 The 'Morphology Gene' Objection

A second, possibly deleterious objection to the present account of language acquisition (and evolution) could, perhaps, be based on Gopnik's (1990a, 1990b) and Gopnik & Crago's (1991) work on individuals with 'Specific Language Impairment' (SLI)—also called 'dysphasics'—from a three-generation family. Her studies suggest that these subjects suffer from a selective impairment of their morphological system, while other parts of their language ability is left intact. The cause for this selective deficit is hypothesized to be an abnormality in a dominant gene associated with the learning of abstract morphology. This, in turn, implies the postulation of a separate grammar component for the learning (and processing) of morphology along the lines suggested in Pinker (1984). Some (e.g., notably Pinker 1991, 1994; Pinker & Bloom, 1990) take Gopnik's results to possibly confirm the existence of an innate UG. Indeed, if the above hypothesis is true, language learning (as suggested above) would only be able to play a minor role in the explanation of language acquisition.

But, other explanations of the impairment data are possible. First, there is considerable controversy concerning the interpretation of Gopnik's data and the proposed extent of the deficit. Vargha-Khadem & Passingham (1991), who also are studying the same family, stress that the impaired individuals in addition to their problems with morphology, suffer from a severe developmental speech disorder, as well as having problems with both phonological repetition and aspects of lexical processing. Apart from

the problems with morphology, "the affected members show deficits in understanding many other types of syntactical structure, such as the reversible passive, postmodified subject, relative clause and embedded forms" (p. 226). Moreover, a closer reflection on the data presented in Gopnik & Crago (1991) points to an alternative, processing based reason for the deficit. Gopnik notes that dysphasics generally appear to learn inflected words as whole, unanalyzed lexical items, as if the latter had no internal morphological structure. In addition, they are very late in developing language. Recall that the late language learners of both second and first language (Newport, 1990) also appear to learn words as unanalyzed wholes. So, the late onset of learning (for whatever reason) might put these individuals in a similar position to normal late learners, preventing them from ever acquiring a 'native' language ability. Of course, other factors, such as, abnormal constraints on memory and processing, are likely to play a part, too. Recent results obtained by Blackwell & Bates (1993) complement this picture. They were able to induce an agrammatic profile of morphological errors in normal adults simply by increasing the subjects' processing load via a secondary task. The results "suggest that this selective profile does not necessarily indicate the existence of a distinct sub-system specialized for the implicated aspects of syntax, but rather may be due to the vulnerability of these forms in the face of global resource diminution" (p. 2). Hence, it seems that a processing based explanation of Gopnik's results (also hinted at by Fletcher, 1990) is possible, which, in turn, takes the sting out of this second objection[21].

Having thus presented my account of language evolution and development, I have, as it turns out, to a certain degree followed a recipe put forward by Tooby & Cosmides (1990): "If one proposes that the ability to acquire a human language is a spandrel of general purpose learning mechanisms, one must state exactly what those general purpose mechanisms are, show that they exist, demonstrate that they are adaptations, and then demonstrate that these general purpose mechanisms can, in fact, allow one to learn language (through, for example, a learnability analysis ...)" (p. 762). Admittedly, the theory outlined in this chapter does not fully meet this challenge. In particular, I have not stated exactly what the general learning mechanisms might be. Rather, I have presented a program for how the challenge might be addressed, proposing to construe language as a nonobligate symbiant evolved through natural selection.

[21] Of course, data concerning other kinds of developmental language impairments (such as, the cases of 'linguistic savants', Smith & Tsimpli, 1991) may serve as the basis for a similar kind of objections. However, space does not allow me to address such possible objections here. I submit my response above to Gopnik's results as a possible template for future rebuttals of such objections.

As such, language may be a spandrel from a hominid perspective, whereas the underlying learning mechanisms are not. The latter are likely to have evolved to permit the extraction of complex sequential information from the environment. The mechanisms subserving language acquisition may therefore not be language-specific, but neither are they totally domain-general either. The present theory thus seems to be incompatible with the four other positions presented in Figure 5.1, perhaps positioning itself somewhere in between them. Of course, there are many details not yet accounted for (as with any other theory). Whether the presented framework will be able to stand the test of time, only future research can tell. The concluding chapter will therefore present some suggestions for future work. Nonetheless, I hope to have shown here that a learning and processing based theory of language is, at least, possible.

Chapter 6

Conclusion

To many, the theory of an innate UG provides the only viable explanation of the acquisition of language in the face of the purported poverty of the stimulus (e.g., Chomsky, 1980, 1986; Crain, 1991; Piattelli-Palmarini, 1989, 1994; Pinker, 1994). Indeed, UG is often characterized as the *'only game in town'*. The idealization of this substantial endowment of language-specific knowledge as the basis of an infinite linguistic competence, to be contrasted with the empirical shortcomings in language performance, has further supported a nativist position. Together, the argument from the poverty of stimulus and the competence/performance distinction have proved to be a major stumbling block for learning and processing based approaches to language. Indeed, it would appear that

> the child's language 'grows in the mind' as the visual system develops the capacity for binocular vision, or as the child undergoes puberty at a certain stage of maturation. Language acquisition is something that happens to a child placed in a certain environment, not something that the child does. (Chomsky, 1993: p. 29).

In this thesis, I challenge the nativist view of the acquisition and processing of language. Language acquisition involves more than the gradual unfolding of a genetic blue-print given minimal environmental stimuli—it involves *learning* (in a non-trivial sense of the latter). I have shown that linguistic competence need not be separated from observable language performance within a connectionist framework and that the primary linguistic stimuli might not be as poor as it first appears. In short, UG is no longer the only game in town. Connectionism has paved the way for an alternative view which eschews the domain-specific nativism of UG, while leaving the door open for innate constraints that are not specific to language.

Chapter 2 showed that the competence/performance distinction is an artifact of construing linguistic grammars as sets of recursive rules. Given the infinite nature of such recursive grammars, a separation is necessary between the latter as an idealized competence and a limited performance in order to account for observable language behavior. Moreover, we saw that the existence of very limited non-iterative recursion does not warrant this distinction. I therefore suggested that the distinction be abandoned for methodological as well as empirical reasons. However, this requires an alternative means of representing the grammatical regularities subserving our language since couching the latter in terms of (recursive) rules is what created the need for the distinction in the first place. It was my contention that we might find such a representational vehicle within connectionism. The simulations reported in chapters 3 and 4 support this assertion, demonstrating that recurrent neural networks are able to capture a limited amount of non-iterative recursion. The graceful degradation of network performance appears to follow human behavior on similar non-iterative recursive structures. In addition, these networks also appear to be capable of the kind of strong generalization we would expect from models of human language learning.

Having shown that connectionist models seem to have sufficient computational power to serve as models of human language behavior, the question still remains of how to overcome the apparent poverty of the stimulus in its many guises. It is clear that the linguistic information available to a child involves both grammatical and ungrammatical utterances. Although this kind of noisy input generally is a problem for classical models of language learning (e.g., Berwick & Weinberg, 1984)—as pointed out in chapter 5—most neural networks models learn the best when faced with such inconsistent input. Another problem addresses the issue of infinite generalization given only finite (positive) input. This problem is solvable, I submit, by connectionist (or other) models incorporating maturational changes (as in the simulations in chapter 4) implemented as decreasing constraints on processing/memory. In this way, early learning may facilitate the acquisition of simple structures which, in turn, can scaffold the subsequent learning of more complex structures. A third problem is the purported inadequacy of empirical learning methods as presented in thought experiments such as 'the competent scientist gambit' (Ramsey & Stich, 1991). This kind of criticism seems targeted at traditional models of learning using induction, hypothesis generation and testing, and so on. However, when it comes to learning driven by subtle statistical contingencies as found in connectionist models, our intuitions fall short, and are even further confused by the incorporation of changing maturational constraints in the learning process. Finally, children very early in their language development appear to

master a number of linguistic principles that are essentially arbitrary and have no direct reflection in the input. It is likely that these particular universals are not learned. Instead, I argue that they are by-products of the learning and processing of language in human infants. As such, the universals are not innately specified chunks of linguistic knowledge, but side-effects of mechanisms that have evolved to learn and process sequential and hierarchical information in general.

This leads us to the topic of the origin and evolution of language, a topic which was banned by the Société de Linguistique de Paris in 1866, but which I nevertheless ventured to discuss in chapter 5. Many evolutionary accounts of language take UG as the end goal of evolutionary processes (e.g., Chomsky, 1988, 1993; Corballis, 1992; Greenfield, 1991; Hurford, 1991; Piattelli-Palmarini, 1989; Pinker, 1994; Pinker & Bloom, 1990). The proponents of exaptationist theories of language evolution propose that UG emerged as a product of genetic hitch-hiking, random mutations, or as a side-effect of increases in the complexity and size of the human brain (following some, yet unknown, natural laws). In contrast, adaptationist UG approaches typically suggest that an increasingly complex language ability was selected for because it provides an efficient means of communication. However, this position is not as easy to defend as it first appears given the assumption of UG. The linguistic principles proposed by the theory of UG manifest themselves as not serving any communicative functions. In fact, they provide arbitrary, idiosyncratic restrictions on the amount of information that can be exchanged. Thus, it seems that "survival criteria, the need to communicate and plan concerted action, cannot account for our *specific* linguistic nature" (Piattelli-Palmarini, 1989: p.25).

It is difficult to explain how language could have evolved through natural selection to its present form given that its universals appear to have no functional explanation in purely linguistic terms. How could more progressively complex grammars have been selected for when the complex grammatical principles serve no communicative function, but often rather impede it? Pinker & Bloom (1990) bravely take on the task of defending the adaptationist UG position, arguing that the linguistic universals function as part of a standardized communication protocol. It therefore does not matter what the universals are *qua* standards as long as everybody in a particular speech community adopts the same set of standards. Pinker & Bloom furthermore stress that language shows evidence of design in much the same way as the vertebrate visual system does. Although this points to an adaptationist explanation, there is an important disanalogy between language and vision. When explaining the evolution of the vertebrate visual system we can point to functional aspects of its parts as having adaptive value, such as, for example, the selective advantage of having an illumination sensitive pupil that

permits the detection of objects given variable light intensity (within certain limits). The universal principles of UG, on the other hand, cannot be explained in the same way since they serve no communicative function.

If UG is assumed to be the end goal of language evolution, then we are forced to entertain either the, presently, rather unsubstantiated ('hopeful monster') position of the exaptationists, or the adaptationist counterpart with its problematic explanation of linguistic universals. But if UG is not an *a priori* assumption, then we might find a more satisfying account of the origin and evolution of language. In chapter 5, I presented an alternative evolutionary perspective on language without recourse to UG, construing language as a non-obligate symbiant adapting to the idiosyncrasies of human learning and processing mechanisms. In this picture, mechanisms for the learning and processing of hierarchical, sequential information preceded language (as evidenced by the former's spread across a vast variety of species). Since one of the main characteristics of language is its temporal and sequential nature, it is likely that language in its origin was subserved by the evolutionarily more ancient sequential learning mechanisms. Language then gradually evolved from being a somewhat limited manual language of gestures into a predominately vocal language. Syntactic complexity presumably increased following decreases in morphological complexity, the latter being a product of fitting of a growing vocabulary within the memory capacities of finite minds.

An important remaining question is whether subsequent evolution may have caused the acquisition of language to become largely innately driven. Pinker & Bloom (1990) suggest that the Baldwin effect (Baldwin, 1896) perhaps allowed such a gradual transformation from a largely learning based language ability to one determined by an innate UG. I have argued that this scenario is not a plausible one given the very nature of language acquisition. One of the premises of the Baldwin effect is that the trait, which is to become innate, should remain constant. This is not true in the case of language acquisition since the rate of linguistic change is considerably faster than the rate of evolutionary change in humans, making language a 'moving target' *vis-a-vis* the Baldwin effect. Recall also that on the present account, cross-linguistic universals are by-products of learning and processing, and are therefore prevented from becoming innate. Moreover, the simulation experiments by French & Messinger (1994) can be taken to indicate that if most hominids were able to learn the early forms of language, then only small changes would be seen in terms of innate specification of learning mechanisms; perhaps, towards faster processing of more complex, sequential information. This, in turn, would permit language to become more complex, and therefore afford better communication, producing a pressure on the hominids to be able to learn and process even more complex structures, and so on.

Summarizing, in this thesis I set out to explain how the ability for infinite languages might fit within finite minds. I have demonstrated that certain kinds of finite-state automata—i.e., recurrent neural networks—are likely to have sufficient computational power and the necessary generalization capability to serve as models for the processing and acquisition of language. An evolutionary picture was furthermore provided, suggesting a learning and processing based explanation of the origin and subsequent phylogenetic development of language. This paved the way for an account of language acquisition, which relies on maturationally constrained learning processes, and challenges the poverty of the stimulus argument. Importantly, my theory of the human language ability, as sketched in this thesis, cuts across both the dichotomy between exaptationist and adaptationist accounts of language evolution and the dichotomy between domain-specific and domain-general language learning. Language has adapted itself to sequential learning and processing mechanisms existing prior to the appearance of language. These mechanisms presumably also underwent changes after the emergence of language, but language was not the only evolutionary pressure causing this change. Other kinds of hierarchical processing, such as, the need for increasingly complex manual combination following tool sophistication, were also contributing to this change. This means that although the mechanisms subserving language are products of adaptation, they were not adapted solely for the purpose of language. Instead, they are products of adaptations for the processing and learning of hierarchically organized sequential information. Language learning in this picture is therefore neither domain-specific with respect to language, nor is it fully domain-general, rather, it is oriented towards the domain of sequential structure. As such, my theory of language processing, acquisition, and evolution makes a number of predictions whose investigation might determine whether the theory will stand the test of time and further scientific scrutiny. I will therefore close this thesis by mentioning a few of these predictions and suggesting ways to research them.

6.1 Future Directions

Language Universals: $\overline{\text{X}}$-theory

One of the remaining and most pressing problems for the theory presented in this thesis is to account for universal linguistic principles other than subjacency within a learning and processing based framework. This is a nontrivial task which is further complicated by the fact that different linguistic theories prescribe different universals. Nonetheless, some universals appear to hold across different linguistic approaches, and

thus provide suitable starting points for such investigations. For example, the $\overline{\text{X}}$-theory of phrase structure (Jackendoff, 1977) is used in both GB (e.g., Chomsky, 1981, 1986) and GPSG (Gazdar *et al.*, 1985). Since it is beyond the scope of this thesis to explain all linguistic universals, I concentrate on the above example and suggest that this preliminary explanation might serve as a template for the further investigation of other universals.

There is a clear statistical tendency across all human languages to conform to a format in which the head of a phrasal category consistently is placed in the same position—either first or last—with respect to the remaining clause material[1]. English is a head-first language; meaning that the head is always placed first in a phrase, as when the verb is placed before the object NP in a transitive VP such as *'eat licorice'*. In contrast, speakers of Hindi would say the equivalent of *'licorice eat'*, because Hindi is a head-last language. More generally, the phrase structure regularities of any language in most cases have the basic structure of either (1) or (2):

(1) $[_{\text{XP}}$ X ...] (e.g., $[_{\text{VP}}$ V ...], $[_{\text{NP}}$ N ...], $[_{\text{PP}}$ prep ...], etc.)
(2) $[_{\text{XP}}$ ... X] (e.g., $[_{\text{VP}}$ ... V], $[_{\text{NP}}$ ... N], $[_{\text{PP}}$ ... post], etc.)

In theory of UG, it is argued that knowledge of $\overline{\text{X}}$-theory is innate (e.g., Chomsky, 1986) and all that remains for the child to learn is whether her language is head-first or head-last.

However, I contend that the general tendency for languages to conform to these ordering principles may be a by-product of certain constraints on language learning. In particular, languages may need to have a certain consistency across their different grammatical regularities in order for the former to be learnable by learning devices with adapted sensitivity to sequential information. Languages that do not have this kind of consistency in their grammatical structure are perhaps not learnable, and they will, furthermore, be difficult to process (cf. Hawkins, 1990). A possible way of demonstrating this point is to train a connectionist network on a language conforming to (1), another on (2), and a third net on a language such as (3), which violates the ordering principles of $\overline{\text{X}}$-theory:

(3) $[_{\text{VP}}$ ...V], $[_{\text{NP}}$ N ...], $[_{\text{NP}}$ ...$\overline{\text{S}}$ N], $[_{\text{PP}}$ prep ...], etc.

We have already seen in chapter 4 that a language with the basic structure of (1) appears to be learnable by recurrent neural networks (even when cross-dependency

[1]It should be noted, however, that *pace* Pinker(1994) all languages do not conform the ordering conventions of $\overline{\text{X}}$-theory. For example, Hawkins (1990) mentions that Mandarin Chinese has both VO and OV word order as well as both prepositions and postpositions. This is an exception which Hawkins attributes to processing.

structures are present). It therefore seems very likely that a language structurally conforming to (2) would also be learnable. In contrast, I predict that a language with the structure of (3) will not be learnable. Whether the results of such simulations will support this prediction, and thus my view of $\overline{\text{X}}$-theory as an artifact of learning and processing constraints, cannot be determined *a priori*. Interestingly, the grounds for some optimism regarding this project can be found in Hawkins' (1990) parsing theory of word order universals, which suggests that constraints on memory and processing may play a considerable role in determining grammatical structure. Although, he still leaves room for innate grammatical knowledge, Hawkins also proposes that "the parser has shaped the grammars of the world's languages, with the result that actual grammaticality distinctions, and not just acceptability intuitions, performance frequencies, and psycholinguistic experimental results ..., are ultimately explained by it" (p. 258). And this is precisely what my theory would predict.

Simulating Creolization

In chapter 5, I hinted at another learning based solution to a problem which has been taken by many (e.g., Bickerton, 1984; Pinker, 1994) to require an innate UG: the creolization of language. Recall that this process allows children to overcome a considerable amount of inconsistency in their linguistic input, as in the case of the creolization of pidgins (Bickerton, 1984), and in Simon's ability to surpass the performance of his parents despite only getting their 'approximated ASL' as input (Singleton & Newport, 1993). Here I focus on the latter, since this study provides a qualitative and more detailed description of the effects of creolization. One of the important questions raised by this research is whether Simon's impressive feat can be explained by processes of learning without recourse to innate linguistic knowledge. Before answering that question it is worthwhile looking closer at Singleton & Newport's findings. They argue that Simon is responding to the inconsistent input from his parents by way of *'frequency boosting'*; that is, he is extracting the most frequent, consistently used forms from the input provided by his parents and boosting the frequency of these forms in his own output. In this process he is not relying entirely on absolute frequency information, but only boosts those forms which show consistency in terms of correlation between form and meaning. These forms, in turn, largely coincide with ASL forms, thus permitting Simon to exhibit a signing behavior very close to the ASL of early learners (with consistent input). In effect, he acquires a (near normal) systematic morphology by statistically filtering out the noise from his parents output.

Returning to the question of how this might be explained in terms of learning, I

in unison with Singleton & Newport suggest that maturationally constrained learning mechanisms underlie the process of creolization. The idea is that such learning mechanisms change their own *effective* input (as mentioned in chapter 5), and may therefore be able to filter out the noise that hide the grammatical structure. There are a number of ways to pursue this idea in terms of simulation experiments. I mention three here. First, a recurrent neural network undergoing maturational change, similar to the nets used in chapter 4, could be trained on a corpus from a small artificial pidgin language (perhaps paired with some minimal contextual information). The outcome of this training is predicted to be a creole which has more grammatical structure than the pidgin. The second proposed simulation involves the learning of morphological structure via a pairing of linguistic form with meaning, extending the simulations reported in Goldowsky & Newport (1993). In the proposed experiment, a maturationally constrained recurrent network is to be trained on artificial data corresponding in statistical structure to the input available to Simon; i.e., with roughly 70% consistent form-meaning pairings, leaving the remaining 30% to reflect the typical inconsistency errors made by his parents. Again, the prediction is that the net may creolize the input, and, in this case, perhaps exhibit frequency boosting similar to what Simon evinced. The last simulation elaborates on the previous experiment by starting with the training of a parent network. However, the parent net will not undergo maturational changes, but has the adult memory capacity throughout its period of learning in order to simulate late learning. This net should furthermore be trained on a fully consistent corpus of form-meaning pairings, signaling the idealization that the parents learned ASL from perfect input. Once that net is fully trained, its output will serve as input for a child net. If the present ideas are right, the parent net may display a typical late learner output—such as the input Simon received from his parents. These proposed simulation experiments are admittedly somewhat speculative, but they receive some credibility from the results already reported here in chapter 4, in Elman (1993), and in Goldowsky & Newport (1993). Moreover, Singleton & Newport (1993) argue that learning mechanisms not specific to language may be able to account for creolization as a statistical process of frequency boosting.

Testing Incremental Memory Learning Experimentally

The role that maturationally constrained learning processes are purported to play in creolization and, more generally, in language acquisition also suggests a possible psycholinguistic experiment. If a child's learning mechanisms undergo maturational changes resembling those simulated in chapter 4 (and in the work of Elman, 1991b,

1993, and Goldowsky & Newport, 1993), we may expect to be able to induce the effects of this kind of constrained learning in adults. The motivation is that if we can impose gradually decreasing memory constraints on adult subjects in an artificial grammar learning task, then these subjects may improve their performance on this task compared with a control group learning the task without additional memory constraints. That is, initial memory constraints may paradoxically facilitate learning by forcing the subject to detect simple correlational structures which, in turn, can be used to scaffold the acquisition of more complex statistical contingencies. In other words, we may be able to induce the effects of Newport's (1990) less-is-more hypothesis in adults.

To investigate this prediction, I propose to submit the subjects to an implicit learning task where they are to learn the regularities underlying a simple context-free grammar involving center-embeddings. It is known that people are able to learn the iterative recursive structure of regular grammars (for an overview, see Reber, 1989), but they do not appear to be able to acquire non-iterative recursive constructs from a context-free grammar (Poletiek, 1994). However, if my prediction is correct, then we may expect a gradually decreasing memory constraint to facilitate the learning of the latter structures. Rather than operationalizing this memory constraint in the form of a secondary distractor task (such as, keeping a set of numbers in memory), I intend to use a functional operationalization of it. In the particular kind of implicit learning task that is relevant here, the subjects are asked to memorize strings from a small artificial grammar. The control group in the proposed experiment will be presented with context-free strings, e.g., *'NVcnNVv'*. The subjects learning under the memory constraint condition will see the same strings as the control group, but this time divided into randomly ordered chunks, such as *'Vv cnN NV'*. These chunks correspond functionally to the effects of looking at the string with a 2-3 word memory window, and will gradually increase in size until the subjects are exposed to a number of complete strings in the last phase of learning. If the constrained memory condition facilitates learning, as predicted, then we have some additional empirical evidence for the less-is-more hypothesis, and thus for the theory of language acquisition outlined in chapter 5.

Simulating the Evolution of Language

Moving to my evolutionary account of language, I argued there that the constancy assumption underlying the Baldwin effect is not likely be met in the case of language, because the latter constitutes a 'moving target'. The Baldwin effect may therefore be an implausible candidate mechanism by which a learning based language ability becomes innate in the sense of UG. I would like to provide additional weight to this

argument by way of simulated evolution. The idea is to simulate the evolution of a learning mechanism which has to adapt itself towards becoming a better learner of an ever changing language. A recent evolutionary simulation, reported in Batali (1994), is a step in this direction. Batali trained a population of simple recurrent networks on a language from a set of 36 languages with different (4 item) vocabularies, but with the same inherent context-free structure (resembling the counting recursive languages used in chapter 3). Networks from different generations faced languages with different vocabularies. Over 'evolutionary time' the networks developed an initial set of weights biased towards the learning of the whole group of languages, rather than a specific language. In fact, the average performance after 150 generations was better than the best performance found in a pool of 436 randomly initialized, and subsequently trained, networks. At first glance, this result seems to suggest that the Baldwin effect may work to make language learning innate. However, a closer analysis points in the opposite direction. Importantly, the same amount of training is needed to reach asymptotic performance for both the evolved and the randomly initialized nets. Thus, learning plays the same important role in the acquisition of language in both cases. What the evolutionary process here has endowed the networks with is learning *biases*, not the kind of innate *knowledge* presupposed in UG[2] (but notice that the set of languages used in this experiment may have been too simple to require innate knowledge for optimal acquisition).

Batali's (1994) experiments do not simulate the evolutionary scenario I sketched in chapter 5. Crucially, language itself did *not* undergo evolutionary change; instead, the difference between the languages presented to the nets was merely a difference in the function (or 'meaning') of a fixed set of vocabulary items, not in language structure *per se*. I therefore propose to conduct simulations in which both learning mechanism and language are subject to evolutionary change. Just as each initial configuration of a learning mechanism will be determined by a genome (specifying the initial weights in a simple recurrent network, size of network, and maturational changes), so will a 'genome' specify the layout of each language (in terms of properties such as, head positioning in phrases, the existence of center-embedding or cross-dependency, and degree of morphological regularization). Allowing both learning mechanism and language to evolve in a dynamical relationship—in which the learning mechanism constrains the language more than *vice versa*—may show that the Baldwin effect cannot function

[2]In addition, if the present theory of language evolution is correct, then the same sequential learning mechanisms that subserve language may also subserve hierarchical manual combination (as suggested by Greenfield, 1991). This means that these learning mechanisms cannot adapt themselves specifically for language, but must adapt to the more general requirements of sequence learning and processing.

when it is 'chasing' a moving target (or, only causes very little of this learning process to become innate). Other avenues for experimentation follow from this set-up; here I mention but one. An evolutionary experiment could be run to test the prediction that the accommodation of a growing vocabulary within a fixed sized network would force an initially morphologically complex language towards a language with a highly regularized morphology. The results obtained by Hare & Elman (1994) indicate that when a net's memory capacity is over-stretched, we may expect a shift towards a more regularized morphology. Computational resources permitting, possible changes in syntax following such morphological regularization could be studied to see if the latter would lead to a more complex syntax in order to maintain ease of communication.

Further Psycholinguistic Experimentation

A number of other experiments and topics for further theoretical research come to mind and which may provide the basis for further investigations of the theory of language presented in this thesis; but time and space do not allow me to treat them in any detail here. In chapter 2, we saw that empirical data suggested a marked difference in performance on iterative and non-iterative recursive structures. This difference was mirrored in the simulation results in chapter 4, however, the latter also suggested certain limitations on the processing of iterative recursion. It would be interesting to study human performance on these structures (perhaps in terms of recall, comprehension, or grammaticality judgement) to see whether it corresponds to the predictions made from network performance. Another prediction following the discussions in chapter 2, is that it may be possible to train people to improve their performance on sentences that are ungrammatical in the same systematic way, just as they via training can improve their performance on center-embedded structures. If this prediction was borne out experimentally, then the arguments against the competence/performance distinction would be further strengthened. In addition to these experimental investigations, a detailed study of acquired and developmental language disorders in the light of the present theory might provide supplementary support for the latter (or, perhaps, suggest areas in need of revision).

In this thesis, I have outlined a theory of the learning, processing, and evolution of language, and suggested how connectionists networks might function as models of our language ability. For reasons of practicality, I have concentrated on linguistic structure, leaving other aspects of language aside, such as, semantics and pragmatics. In doing so, I have made the idealization that syntax can be treated independently of these other

aspects. However, I am not convinced that such a separation is warranted, or even useful in the long run, and in invoking this separation I might have fallen in the incommensurability trap. Still, I had to start somewhere, and linguistic structure seems to be as good as place as any—especially, given the amount of empirical data that have been elicited concerning syntactic processing. Moreover, the earlier mentioned simulations by Weckerly & Elman (1992) did suggest that (minimal) semantic information can readily be incorporated into this kind of model, a point which has been further corroborated by St. John & McClelland's (1990) model of the learning and application of contextual constraints in sentence processing (though they use a slightly more complex network configuration). So, despite the omission of semantic and contextual considerations, I believe that the present theory permits yet another step towards a learning and processing based explanation of the acquisition and evolution of language. And this may, perhaps, lead us closer to solving the age old question of how finite minds create infinite languages.

References

Abrams, M. & Reber, A.S. (1988) Implicit Learning: Robustness in the Face of Psychiatric Disorders. *Journal of Psycholinguistic Research*, **17**, 425–439.

Altmann, G. & Steedman, M. (1988) Interaction with Context in Human Sentence Processing. *Cognition*, **30**, 191–238.

Bach, E., Brown, C. & Marslen-Wilson, W. (1986) Crossed and Nested Dependencies in German and Dutch: A Psycholinguistic Study. *Language and Cognitive Processes*, **1**, 249–262.

Baldwin, J.M. (1896) A New Factor in Evolution. *American Naturalist*, **30**, 441-451.

Barwise, J. (1989) Unburdening the Language of Thought. Chapter 7 in *The Situation in Logic*. Stanford, CA: CSLI.

Batali, J. (1994) Artificial Evolution of Syntactic Aptitude. In *Proceedings from the Sixteenth Annual Conference of the Cognitive Science Society*, 27–32. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bates, E.A. & Elman, J.L. (1993) Connectionism and the Study of Change. In M.J. Johnson (Ed.), *Brain Development and Cognition*. Cambridge, Mass.: Basic Blackwell.

Bates, E.A., Thal, D. & Marchman, V. (1991) Symbols and Syntax: A Darwinian Approach to Language Development. In N.A. Krasnegor, D.M. Rumbaugh, R.L. Schiefelbusch & M. Studdert-Kennedy (Eds.), *Biological and Behavioral Determinants of Language Development*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bechtel, W. (1989) Connectionism and Intentionality. In *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bechtel, W. & Abrahamsen, A. (1991) *Connectionism and the Mind: An Introduction to Parallel Distributed Processing in Networks*. Oxford: Basil Blackwell.

Berwick, R.C & Weinberg, A.S (1984) *The Grammatical Basis of Linguistic Performance: Language Use and Acquisition*. Cambridge, MA: MIT Press.

Bickerton, D. (1984) The Language BioProgram Hypothesis. *Behavioral and Brain Sciences*, **7**, 173–212.

Blackwell, A. & Bates, E. (1993) Inducing Agrammatic Profiles in Normals: Evidence for the Selective Vulnerability of Morphology under Cognitive Resource limitation. Technical Report, No. CRL-9304, Centre for Research in Language, University of California, San Diego.

Blaubergs, M.S. & Braine, M.D.S. (1974) Short-term Memory Limitations on Decoding Self-embedded Sentences. *Journal of Experimental Psychology*, **102**, 745–748.

Bloom, P. (1994) Generativity within language and other cognitive domains. *Cognition*, **51**, 177–189.

Brill, E. Magerman, D. Marcus, M. & Santorini, B. (1990) Deducing Linguistic Structure from the Statistics of Large Corpora. In *DARPA Speech and Natural Language Workshop*. Hidden Valley, Pennsylvania: Morgan Kaufmann.

Bruce, V. & Green, P. (1985) *Visual Perception, Physiology, Psychology, and Ecology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Capaldi, E.J. & Miller, D.J. (1988) The Rat's Simultaneous Anticipation of Remote Events and Current Events Can be Sustained by Memories Alone. *Animal Learning and Behavior*, **16**, 1–7.

Chalmers, D.J. (1990a) Syntactic transformations on Distributed Representations. *Connection Science*, **2**, 53–62.

Chalmers, D.J. (1990b) Why Fodor and Pylyshyn were Wrong: The Simplest Refutation. In *Proceedings of the Twelveth Annual Meeting of the Cognitive Science Society*, 340–347. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chater, N. (1989) Learning to Respond to Structures in Time. Technical Report, Research Initiative in Pattern Recognition, St Andrews Road, Malvern, Worcs. RIPRREP/1000/62/89.

Chater, N. & Conkey, P. (1992) Finding Linguistic Structure with Recurrent Neural Networks. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chater, N., & Oaksford, M.R. (1990) Autonomy, Implementation and Cognitive Architecture: A Reply to Fodor and Pylyshyn. *Cognition*, **34**, 93–107.

Chomsky, N. (1956) Three Models for the Description of Language. *I.R.E. Transactions on Information Theory*, **IT-2**, 113-124.

Chomsky, N. (1957) *Syntactic Structures*. The Haque: Mouton.

Chomsky, N. (1959a) On certain Formal Properties of Grammars. *Information and Control*, **2**, 137–167.

Chomsky, N. (1959b) Review of Skinner (1957). *Language*, **35**, 26–58.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, Mass: MIT Press.

Chomsky, N. (1972) *Language and Mind*. Harcourt, Brace and World (extended edition).

Chomsky, N. (1975) *Reflections on Language*. New York: Pantheon.

Chomsky, N. (1976) On the Nature of Language. *Annals of the New York Academy of Sciences*, **280**, 46–55.

Chomsky, N. (1977) *Language and Responsibility*. New York: Pantheon.

Chomsky, N. (1980). *Rules and Representations* New York: Columbia University Press.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Forris Publications.

Chomsky, N. (1986). *Knowledge of Language*. New York: Praeger.

Chomsky, N. (1988) *Language and the Problems of Knowledge. The Managua Lectures*. Cambridge, Mass: MIT Press.

Chomsky, N. (1993) *Language and Thought*. Wakefield, RI: Moyer Bell.

Christiansen, M. (1992) The (Non)Necessity of Recursion in Natural Language Processing. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Christiansen, M. & Chater, N. (1992) Connectionism, Learning and Meaning. *Connection Science*, **4**, 227–257.

Christiansen, M. & Chater, N. (1994) Generalization and Connectionist Language Learning. *Mind and Language*, **9**.

Church, K. (1982) *On Memory Limitations in Natural Language Processing*. Bloomington, IN: Indiana University Linguistics Club.

Clark, A. (1993) Minimal Rationalism. *Mind*, **102**, 587–610.

Cleeremans, A. (1993) *Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing*. Cambridge, Mass.: MIT Press.

Cleeremans, A., Servan-Schreiber, D. & McClelland, J. L. (1989) Finite state automata and simple recurrent networks. *Neural Computation*, **1**, 372–381.

Cliff, D. T. (1990) Computational Neuroethology: A Provisional Manifesto. Techical Report No. CSRP-162, School of Cognitive and Computing Sciences, University of Sussex, Brighton.

Corballis, M.C. (1992) On the Evolution of Language and Generativity. *Cognition*, **44**, 197–226.

Corballis, M.C. (1994) The Generation of Generativity: a response to Bloom. *Cognition*, **51**, 191–198.

Cottrell, G. W. & Plunkett, K. (1991) Learning the past tense in a recurrent network: Aquiring the mapping from meanings to sounds. In *Proceedings of the Thirteenth Annual Meeting of the Cognitive Science Society*, 328–333. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crain, S. (1991) Language Acquisition in the Absence of Experience. *Behavioral and Brain Sciences*, **14**, 597–650.

Crain, S. & Steedman, M. (1985) On Not Being Led up the Garden Path: The Use of Context by the Psychological Syntax Processor. In D.R. Dowty, L. Karttunen & A.M. Zwicky (Eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*. Cambridge: Cambridge University Press.

Darwin, C. (1900) *The Descent of Man, and Selection in Relation to Sex* (2nd Edition). New York: P.F. Collier and Son.

Davies, M. (1993) Two Notions of Implicit Rules. Ms. Oxford University.

Dennett, D.C. (1991) Mother Nature Versus the Walking Encyclopedia: A Western Drama. In W. Ramsey, S. Stich, & D. Rummelhart (Eds.), *Philosophy and Connectionist Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

De Roeck, A., Johnson, R., King, M., Rosner, M., Sampson, G. & Varile, N. (1982) A Myth about Centre-embedding. *Lingua*, **58**, 327–340.

Durkin, K. (1989) Implicit Memory and Language Acquisition. In S. Lewandowsky, J.C. Dunn & K. Kirsner (Eds.), *Implicit Memory: Theoretical Issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ejerhed, E. (1982). The Processing of Unbounded Dependencies in Swedish. In E. Engdahl & E. Ejerhed (Eds.), *Readings on Unbounded Dependencies in Scandinavian Languages*. Stockholm: Almqvist & Wiksell International.

Ellis, A.W. & Young, A.W. (1988) *Human Cognitive Neuropsychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Elman, J. L. (1988) Finding structure in time. Technical Report, No. CRL-8801, Centre for Research in Language, University of California, San Diego.

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

Elman, J. L. (1991a) Distributed Representation, Simple Recurrent Networks, and Grammatical Structure. *Machine Learning*, **7**, 195–225.

Elman, J. L. (1991b) Incremental Learning, or The Importance of Starting Small. In *Proceedings from the Thirteenth Annual Conference of the Cognitive Science Society*, 443–448. Hillsdale, NJ: Lawrence Erlbaum Associates.

Elman, J. L. (1992) Grammatical Structure and Distributed Representations. In S. Davis (Ed.), *Connectionism: Theory and Practice*. New York: Oxford University Press.

Elman, J. L. (1993) Learning and Development in Neural Networks: The Importance of Starting Small. *Cognition*, **48**, 71–99.

Ferreira, F. & Clifton, C. (1986) The Independence of Syntactic Processing. *Journal of Memory and Language*, **25**, 348–368.

Finch, S. & Chater, N. (1992) Bootstraping Syntactic Categories by Unsupervised Learning. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Finch, S. & Chater, N. (1993) Learning Syntactic Categories: A Statistical Approach. In M. Oaksford & G.D.A. Brown (Eds.), *Neurodynamics and Psychology.* New York: Academic Press.

Fletcher, P. (1990) Speech and Language Defects. *Nature*, **346**, 226.

Fodor, J. A. (1975) *The Language of Thought.* New York: Thomas Crowell.

Fodor, J. A. & McLaughlin, B. P. (1990) Connectionism and the Problem of Systematicity. Why Smolensky's Solution Doesn't Work. *Cognition*, **35**, 183–204.

Fodor, J.A., & Pylyshyn, Z.W. (1988) Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, **28**, 3–71.

Foss, D.J. & H.S. Cairns (1970) Some Effect of Memory Limitations upon Sentence Comprehension and Recall. *Journal of Verbal Learning and Verbal Behavior*, **9**, 541–547.

Frazier, L. & Fodor, J.D. (1978). The Sausage Machine: A New Two Stage Parsing Model. *Cognition*, **6**, 291-325.

French, R.M. & Messinger, A. (1994) Genes, Phenes and the Baldwin Effect: Learning and Evolution in a Simulated Population. In R. Brooks & P. Maes (Eds.), *Artificial Life IV.* Cambridge, MA: MIT Press.

Gazdar, G. & Pullum, G.K. (1985) Computationally Relevant Properties of natural Languages and Their Grammars. Technical Report CSLI-85-24. Center for the Study of Language and Information, Stanford University.

Gazdar, G., Klein, E., Pullum, G. & Sag, I. (1985). *Generalized Phrase Structure Grammar.* Oxford: Basil Blackwell.

George, A. (1989) How Not to Become Confused about Linguistics. In A. George (Ed.), *Reflections on Chomsky.* Cambridge, MA: Basil Blackwell.

Gold, E. (1967) Language Identification in the Limit. *Information and Control*, **16**, 447–474.

Goldowsky, B.N. & Newport, E.L. (1993) Modeling the Effects of Processing Limitations on the Acquisition of Morphology: The Less is More Hypothesis. In E. Clark (Ed.), *The Proceedings of the Twenty-fourth Annual Child Language Research Forum.* Stanford, CA: CSLI.

Gopnik, M. (1990a) Feature-blind Grammar and Dysphasia. *Science*, **344**, 715.

Gopnik, M. (1990b) Genetic Basis of Grammar Defect. *Science*, **347**, .

Gopnik, M. & Crago, M.B. (1991) Familial Aggregation of a Developmental Language Disorder. *Cognition*, **39**, 1–50.

Gould, S.J. (1979) Panselectionist Pitfalls in Parker & Gibson's Model of the Evolution of Intelligence. *Behavioral and Brain Sciences*, **2**, 385–386.

Gould, S.J. & Lewontin, R.C. (1979) The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London*, **205**, 281–288.

Gould, S.J. & Vrba, E.S. (1982) Exaptation - A Missing Term in the Science of Form. *Paleobiology*, **8**, 4–15.

Gould, S.J. (1993) *Eight Little Piggies: Reflections in Natural History.* New York: Norton.

Greenfield, P.M. (1991) Language, Tools and Brain: The Ontogeny and Phylogeny of Hierarchically Organized Sequential Behavior. *Behavioral and Brain Science*, **14**, 531–595.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R. and Wilson, R. (1989) The Learnability and Acquisition of of the Dative Alternation in English. *Language*, 65, 203–257.

Grosjean, F. (1980) Spoken Word Recognition Processes and the Gating Paradigm. *Perception and Psychophysics*, **28**, 267–283.

Grossman, M. (1980) A Central Processor Hierarchically Structured Material: Evidence from Broca's Aphasia. *Neuropsychologia*, **18**, 299-308.

Hadley, R.F. (1993a) The 'Explicit-Implicit' Distinction. Technical Report CSS-IS TR93-02. Centre for Systems Science, Simon Fraser University.

Hadley, R.F. (1993b) Connectionism, Explicit Rules, and Symbolic Manipulation. *Mind and Machines*, **3**, 183–200.

Hadley, R.F. (1994a) Systematicity in Connectionist Language Learning. *Mind and Language*, **9**.

Hadley, R.F. (1994b) Systematicity Revisited: Reply to Christiansen & Chater and Niklasson & van Gelder. *Mind and Language*,

Hale, K., Krauss, M., Watahomigie, L., Yamamoto, A., Craig, C., Jeanne, L. & England, N. (1992) Endangered Languages. *Language*, **68**, 1–42.

Hanson, S.J. & Kegl, J. (1987) PARSNIP: A Connectionist Network that Learns Natural Language Grammar from Exposure to Natural Language Sentences. In *Proceedings of the Eight Annual Meeting of the Cognitive Science Society*, 106–119. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hare, M. & Elman, J.L. (1994) Learning and Morphological Change. Unpublished Ms. Centre for Research in Language, University of California, San Diego.

Harris, C.L. (1991) Alternatives to Linguistic Arbitrariness. *Behavioral and Brain Sciences*, **14**, 622–623.

Hatfield, G. (1991) Representation in Perception and Cognition: Connectionist Affordances. In W. Ramsey, S. Stich, & D. Rummelhart (Eds.), *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hauser, M.D. (1991) If You've Got it, Why not Flaunt it? Monkeys with Broca's Area but no Syntactical Structure to their Vocal Utterances. *Behavioral and Brain Science*, **14**, 564.

Hinton, G.E. & Nowlan, S.J. (1987) How Learning Can Guide Evolution. *Complex Systems*, **1**, 495–502.

Horrocks, G. (1987). *Generative Grammar*. London: Longman.

Hurford, J.R. (1991) The Evolution of the Critical Period for Language Learning. *Cognition*, **40**, 159–201.

Hudson, R. (1990) *English Word Grammar*. Oxford: Basil Blackwell.

Jacobs, B. (1991) Neurobiology and Language Acquisition: Continuity and Identity. *Behavioral and Brain Sciences*, **14**, 565.

Jelinek, F., Lafferty, J.D. & Mercer, R.L. (1990) Basic Methods of Probabilistic Context Free Grammars. Technical Report RC 16374 (72684), IBM, Yorktown Heights, New York.

Jordan, M. (1986) Serial order: a parallel distributed approach. Institute for Cognitive Science Report, 8604, University of California, San Diego.

Juliano, C. & Tanenhaus, M. (1993) Contingent Frequency Effect in Syntactic Ambiguity Resolution. In *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kaplan, R.M. & Bresnan, J. (1982) Lexical-functional Grammar: A Formal System for Grammatical Representation. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press.

Karmiloff-Smith, A. (1992) Nature, Nuture and PDP: Preposterous Developmental Postulates? *Connection Science*, **4**, 253–270.

King, J. & Just, M.A. (1991) Individual Differences in Syntactic Processing: The Role of Working Memory. *Journal of Memory and Language*, **30**, 580–602.

Kimball, J. (1973). Seven Principles of Surface Structure Parsing in Natural Language. *Cognition*, **2**, 15–47.

Kiparsky, P. (1976) Historical Linguistics and the Origin of Language. *Annals of the New York Academy of Sciences*, **280**, 97–103.

Kirsh, D. (1990) When is Information Explicitly Represented? In P. Hanson (Ed.), *Information, Language and Cognition*, Vancouver, BC: University of British Columbia Press.

Kirsh, D. (1992) PDP Learnability and Innate Knowledge of Language. In S. Davis (Ed.), *Connectionism: Theory and Practice*. New York: Oxford University Press.

Kolb, B. & Whishaw, I.Q. (1990) *Fundamentals of Human Neuropsychology, 3rd edition*. New York: W.H. Freeman and Company.

Kuhl, P.K. (1987) The Special-Mechanisms Debate in Speech Research: categorization Tests on Animals and Infants. In S. Harnad (Ed.), *Categorial Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Kuhl, P.K. (1991) Human Adults and Human Infants Show a "Perceptual Magnet Effect" for the Prototypes of Speech Categories, Monkeys Do Not. *Perception & Psychophysics*, **50**, 93–107.

Kuhl, P.K. (1993) Infant Speech: A Window on Psycholinguistic Development. *International Journal of Psycholinguistics*, **9**, 33–56.

Kuhl, P.K., Williams, K.A, Lacerda, F., Stevens, K.N. & Lindblom, B. (1992) Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. *Science*, **255**, 606–608.

Kuhn, T.S. (1970) *The Structure of Scientific Revolutions* (2nd Edition). Chicago: University of Chicago Press.

Langacker, R.W. (1987) *Foundations of Cognitive Grammar: Theoretical Perspectives. Vol. 1.* Stanford: Stanford University Press.

Larkin, W. & Burns, D. (1977) Sentence Comprehension and Memory for Embedded Structure. *Memory & Cognition*, **5**, 17–22.

Liberman, A.M. & I.G. Mattingly (1989) A Specialization for Speech Perception. *Science*, **243**, 489–494.

Lieberman, P. (1973) On the Evolution of Language: A Unified View. *Cognition*, **2**, 59–94.

Lieberman, P. (1976) Interactive Models for Evolution: Neural Mechanisms, Anatomy, and Behavior. *Annals of the New York Academy of Sciences*, **280**, 660–672.

Lieberman, P. (1991) Speech and Brain Evolution. *Behavioral and Brain Science*, **14**, 566–568.

McClelland, J. L. and Kawamoto, A. H. (1986) Mechanisms of sentence processing. Chapter 19 in J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing*, Volume 2. Cambridge, Mass.: MIT Press.

McClelland, J.L. & Rumelhart, D.E. (Eds.) (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 2: Psychological and Biological Models. Cambridge, Mass.: MIT Press.

MacLennan, B. (1991) Characteristics of Connectionist Knowledge Representation. Technical Report CS-91-147. Computer Science Department, University of Tennessee, Knoxville.

MacWhinney, B. (1993) The (il)Logical Problem of Language Acquisition. In *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. & Snow, C. (1985) The Child Language Data Exchange System. *Journal of Child Language*, 12, 271–295.

Marcus, M. (1980) *A Theory of Syntactic Recognition for Natural Language.* Cambridge, MA: MIT Press.

Marcus, M. (1978) A Computational Account of Some Constraints on Language. In *Theoretical Issues in Natural Language Processing - 2.* New York: Association for Computing Machinery.

Marks, L.E. (1968) Scaling of Grammaticalness of Self-embedded English Sentences. *Journal of Verbal Learning and Verbal Behavior*, **7**, 965–967.

Maskara, A. & Noetzel, A. (1992) Forcing Simple Recurrent Networks to Encode Context. In *Proceedings of the 1992 Long Island Conference on Artificial Intelligence and Computer Graphics.*

Maskara, A. & Noetzel, A. (1993) Sequence Recognition with Recurrent Neural Networks. *Connection Science*, **5**, 139–152.

Matsuzawa, T. (1991) Nesting Cups and metatools in Chimpanzees. *Behavioral and Brain Sciences*, **14**, 570–571.

Maynard-Smith, J. (1978) Optimization Theory in Evolution. *Annual Review of Ecology and Systematics*, **9**, 31–56.

Maynard-Smith, J. (1987) When Learning Guides Evolution. *Nature*, **329**, 761–762.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J. & Amiel-Tison, C. (1988) *Cognition*, **29**, 143-178.

Miller, G.A. (1962) Some Psychological Studies of Grammar. *American Psychologist*, **17**, 748–762.

Miller, G.A. & Chomsky, N. (1963) Finitary Models of Language Users. In R.D. Luce, R.R. Bush & E. Galanter (Eds.), *Handbook of Mathematical Psychology, Vol. II.* New York: Wiley.

Miller, G.A. & Isard, S. (1964) Free Recall of Self-embedded English Sentences. *Information and Control*, **7**, 292–303.

Milne, R.W. (1982) Predicting Garden Path Sentences. *Cognitive Science*, **6**, 349–373.

Narayanan, A. (1992) Is Connectionism Compatible with Rationalism? *Connection Science*, **4**, 271–292.

Neville, H.J. (1993) Neuroiology of Cognitive and Language Processing: Effects of Erly Experience. In M.J. Johnson (Ed.), *Brain Development and Cognition.* Cambridge, Mass.: Basic Blackwell.

Newell, A. and Simon, H.A. (1976) Computer science as empirical inquiry. *Communications of the ACM*, **19**, 113–126. Reprinted in M. Boden (Ed.) *The Philosophy of Artificial Intelligence.* Oxford: Oxford University Press.

Newport, E. (1990) Maturational Constraints on Language Learning. *Cognitive Science*, **14**, 11–28.

Niklasson, L. & Sharkey, N.E. (1992) Connectionism and the Issues of Compositionality and Systematicity. Presented at the 1992 EMCSR Symposium on Connectionism and Cognitive Processing, University of Vienna, April.

Niklasson, L. & van Gelder (1994) Systematicity in Connectionist Language Learning. *Mind and Language*, **9**,

Norris, D. G. (1990) A dynamic-net model of human speech recognition. In G. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and cognitive perspectives.* Cambridge, Mass.: MIT Press.

Oaksford, M., Chater, N., & Stenning, K. (1990) Connectionism, Classical Cognitive Science and Experimental Psychology. *AI and Society*, **4**, 73–90.

Perfetti, C.A. (1990) The Cooperative Language Processors: Semantic Influences in an Automatic Syntax. In D.A. Balota, G.B. Flores d'Arcais & K. Rayner (Eds.), *Comprehension Processes in Reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Piattelli–Palmerini, M. (1989) Evolution, Selection and Cognition: From "Learning" to Parameter Setting in Biology and in the Study of Language. *Cognition*, **31**, 1–44.

Piattelli–Palmerini, M. (1994) Ever since Language and Learning: Afterthoughts on the Piaget-Chomsky Debate. *Cognition*, **50**, 315–346.

Pickering, M. & Chater, N. (1992). Processing Constraints on Grammar. Ms. University of Edinburgh.

Pinker, S. (1984) *Language Learnability and Language Development.* Cambridge, Mass: Harvard University Press.

Pinker, S. (1989) *Learnability and Cognition: The Acquisition of Argument Structure.* Cambridge, Mass: MIT Press.

Pinker, S. (1991) Rules of Language. *Science*, **253**, 530–535.

Pinker, S. (1994) *The Language Instinct: How the Mind Creates Language.* New York: NY: William Morrow and Company.

Pinker, S. & Bloom, P. (1990) Natural Language and Natural Selection. *Behavioral and Brain Sciences*, **13**, 707–784.

Pinker, S. & Prince, A. (1988) On Language and Connectionism. *Cognition*, **28**, 73-195.

Pinker, S., Lebeaux, D.S. & L.A. Frost (1987) Productivity and Constraints in the Acquisition of the Passive. *Cognition*, 26, 195–267.

Plunkett, K. & Marchman, V. (1993) From Rote Learning to System Building. *Cognition*, **48**, 21–69.

Poletiek, F.H. (1994) Implicit Learning of an Artificial Context-free grammar. Paper presented at the Workshop Cognitive Models of language Acquisition, April 21-23, Tilburg, The Netherlands.

Pollack, J.B. (1988) Recursive Auto-Associative Memory: Devising Compositional Distributed Representations. In *Proceedings of the Tenth Annual Meeting of the Cognitive Science Society*, 33–39. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pollack, J.B. (1990) Recursive Distributed Representations. *Artificial Intelligence*, **46**, 77–105.

Port, R. & van Gelder, T. (1991). Representing Aspects of Language. In *Proceedings of the 13th Meeting of the Cognitive Science Society*, 487–492. Hillsdale, NJ: Lawrence Erlbaum Associates.

Prince, A. & Pinker, S (1989) Rules and Connections in Human Language. *Transactions in the Neuro-Sciences.*

Pullum, G.K. & Gazdar, G. (1982) Natural Languages and Context-free Languages. *Linguistics and Philosophy*, **4**, 471–504.

Pulman, S.G. (1986). Grammars, Parsers, and Memory Limitations. *Language and Cognitive Processes*, **2**, 197–225.

Quartz, S.R. (1993) Neural Networks, Nativism, and the Plausibility of Constructivism. *Cognition*, **48**, 223–242.

Quine, W. (1960) *Word and Object.* Cambridge, Mass.: MIT Press.

Ramachandran, V.S. (1993) Behavioral and Magnetoencephalographic Correlates of Plasticity in the Adult Human Brain. In *Proceedings of the National Academy of Science*, **90**, 10413–10420.

Ramsey, W. & Stich, S. (1991) Connectionism and Three Levels of Nativism. In W. Ramsey, S. Stich & D. Rumelhart (Eds.), *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Rayner, K., Carlson, M. & Frazier, L. (1983) The Interaction of Syntax and Semantics during Sentence Processing: Eye Movements in the Analysis of Semantically Biased Sentences. *Journal of Verbal Learning and Verbal Behavior*, **22**, 358–374.

Reber, A.S. (1989) Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology: General*, **118**, 219–235.

Reber, A.S. (1990) On the Primay of the Implicit: Comment on Perruchet and Pacteau. *Journal of Experimental Psychology: General*, **119**, 340–342.

Reber, A.S. (1992) An Evolutionary Context for the Cognitive Unconscious. *Philosophical Psychology*, **5**, 33–51.

Recanzone, G.H. & Merzenich, M.M. (1993) Functional Plasticity in the Cerebral Cortex: Mechanisms of Improved Perceptual Abilities and Skill Acquisition. *Concepts in Neuroscience*, **4**, 1–23.

Redington, M., Chater, N. & M., Finch, S. (1993) Distributional Information and the Acquisition of Linguistic Categories: A Statistical Approach. In *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Reich, P. (1969) The Finiteness of Natural Language. *Language*, **45**, 831–843.

Rumelhart, D.E. & McClelland, J.L.(1986) On the Learning of the Past Tenses of English Verbs. In J.L. McClelland, D.E. Rumelhart and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models.* Cambridge, Mass.: MIT Press.

Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations.* Cambridge, Mass.: MIT Press.

Rumelhart, D.E., McClelland, J.L. & Williams, R.J. (1986) Learning Representations by back-propagating errors. *Nature*, **323**, 533–536.

Schneider, W. (1987) Connectionism: Is it a Paradigm Shift for Psychology? *Behaviour, Research Methods, Instruments, & Computers*, **19**, 73–83.

Seidenberg, M.S. (1994) Language and Connectionism: The Developing Interface. *Cognition*, **50**, 385–401.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J. L. (1989) Learning Sequential Structure in Simple Recurrent Networks in D. Touretsky (Ed.), *Advances in Neural Information Processing Systems*, Vol 1, Morgan Kaufman, Palo Alto, 643–653.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J. L. (1991) Graded State Machines: The Representation of Temporal Contingencies in Simple Recurrent Networks. *Machine Learning*, **7**, 161–193.

Sharkey, N.E. (1991) Connectionist Representation Techniques. *AI Review*, **5**, 143–167.

Sharkey, N.E. (1992) Functional Compositionality and Soft Preference Rules. In B. Linggard & C. Nightingale (Eds.), *Neural Networks for Images, Speech, and Natural Language*. London: Chapman & Hall.

Shieber, S. (1985) *Linguistics and Philosophy*, **8**, 333–343.

Shillcock, R., Levy, J. & Chater, N. (1991) A connectionist model of word recognition in continuous speech. In *Proceedings from the Thirteenth Annual Conference of the Cognitive Science Society*, 340–345. Hillsdale, NJ: Lawrence Erlbaum Associates.

Singleton, J.L. & Newport, E.L. (1993) When Learners Surpass their Models: The Acquisition of American Sign Language from Impoverished Input. Ms. Department of Psychology, University of Illinois at Urbana-Champaign.

Skinner, B.F. (1957) *Verbal Behavior*. New York: Appleton-Century-Crofts.

Slobin, D.I. (1991) Can Crain Constrain the Constraints? *Behavioral and Brain Sciences*, **14**, 633–634.

Smith, N. & Tsimpli, I. (1991) Linguistic Modularity? A Case Study of a 'Savant' Linguist. *Lingua*, **84**, 103–139.

Smolensky, P. (1987) The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, **26**, 137–159.

Smolensky, P. (1988) On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences*. **11**, 1–74.

Sopena, J.M. (1991) ERSP: A Distributed Connectionist Parser that Uses Embedded Sequences to Represent Structure. Technical Report No. UB-PB-1-91. Departament de Psicologia Bàsica, Universitat de Barcelona, Barcelona.

Spivey-Knowlton, M. & Tanenhaus, M. (1994) Immediate Effects of Discourse and Semantic Content in Syntactic Processing: Evidence from Eye-Tracking.

In *Proceedings from the Sixteenth Annual Conference of the Cognitive Science Society*, 812–817. Hillsdale, NJ: Lawrence Erlbaum Associates.

St. John, M.F. & McClelland, J.L. (1990) Learning and Applying Contextual Constraints in Sentence Comprehension. *Artificial Intelligence*, **46**, 217–257.

Stabler, E. (1983) How are Grammars Represented? *Behavioral and Brain Sciences*, **6**, 391–402.

Steedman, M. (1987) Combinatory Grammars and Parasitic Gaps. *Natural Language and Linguistic Theory*, **5**, 403–439.

Stolcke, A. (1991) Syntactic Category Formation with Vector Space Grammars. In *Proceedings from the Thirteenth Annual Conference of the Cognitive Science Society*, 908–912. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stolz, W.S. (1967) A Study of the Ability to Decode Grammatically Novel Sentences. *Journal of Verbal Learning and Verbal Behavior*, **6**, 867–873.

Taraban, R. & McClelland, J.L. (1990) Parsing and Comprehension: A Multiple-constraint View. In D.A. Balota, G.B. Flores d'Arcais & K. Rayner (Eds.), *Comprehension Processes in Reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tooby, J. & Cosmides, L. (1990) Toward an Adaptationist Psycholinguistics. *Behavioral and Brain Sciences*, **13**, 760–762.

van Gelder, T. (1990a) Compositionality: A Connectionist Variation on a Classical Theme. *Cognitive Science*, **14**, 355–384.

van Gelder, T. (1990b) Connectionism and Language Processing. In G. Dorffner (ed.) *Konnektionismus in Artificial Intelligence und Kognitionsforschung*. Berlin: Springer-Verlag.

van Gelder, T. (1991a) Classical Questions, Radical Answers: Connectionism and the Structure of Mental Representations. In T. Horgan & J. Tienson (Eds.) *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer.

van Gelder, T. (1991b) What is the "D" in "PDP"? A Survey of the Concept of Distribution. In W. Ramsey, S. Stich, & D. Rummelhart (Eds.), *Philosophy and Connectionist Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Vargha-Khadem, F. & Passingham, R.E. (1990) Speech and Language Defects. *Nature*, **346**, 226.

Wang, W. S-Y. (1976) Language Change. *Annals of the New York Academy of Sciences*, **280**, 61–72.

Wanner, E. (1980) The ATN and the Sausage Machine: Which One is Baloney? *Cognition*, **8**, 209–225.

Wasow, T. (1991) Debatable Constraints. *Behavioral and Brain Sciences*, **14**, 636–637.

Weckerly, J. & Elman, J. (1992) A PDP Approach to Processing Center-Embedded Sentences. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates.