CONNECTIONIST PSYCHOLINGUISTICS

Morten H. Christiansen, Nick Chater

Greenwood
PUBLISHING GROUP

CONNECTIONIST PSYCHOLINGUISTICS

This page intentionally left blank

CONNECTIONIST PSYCHOLINGUISTICS

Edited by
Morten H. Christiansen
and Nick Chater

ABLEX PUBLISHING
Westport, Connecticut • London

Library of Congress Cataloging-in-Publication Data

Connectionist psycholinguistics / edited by Morten H. Christiansen and Nick Chater.
p. cm.
Includes bibliographical references and index.
ISBN 1-56750-594-5 (alk. paper)—ISBN 1-56750-595-3 (pbk. : alk. paper)
1. Psycholinguistics. 2. Connectionism. I. Christiansen, Morten H., 1963- II. Chater, Nick.
BF455.C677 2001
401'.9—dc21 00-052555

British Library Cataloguing in Publication Data is available.

Library of Congress Catalog Card Number: 00-052555
ISBN: 1-56750-594-5 (hc.)
1-56750-595-3 (pbk.)

First published in 2001

Ablex Publishing, 88 Post Road West, Westport, CT 06881
An imprint of Greenwood Publishing Group, Inc.
www.greenwood.com

Printed in the United States of America

The paper used in this book complies with the Permanent Paper Standard issued by the National Information Standards Organization (Z39.48-1984).

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Connectionist modeling has had a vast impact throughout cognitive science, and has been both highly productive and highly controversial in the area of natural language processing and acquisition. The decade and a half after the publication of David Rumelhart and Jay McClelland's seminal *Parallel Distributed Processing* volumes has seen an explosive growth of connectionist modeling of natural language. During this period the field has matured and is moving away from abstract "existence-proof" models toward making close contact with a range of psycholinguistic data. This book offers the first comprehensive treatment of this emergent area of research, demonstrating the current state of the art (Part I) and appraising the prospects for future development (Part II) of "connectionist psycholinguistics."

The book is based on a special issue of *Cognitive Science*, "Connectionist models of human language processing: Progress and prospects," Vol. 23, no. 4, edited by Morten H. Christiansen, Nick Chater, and Mark S. Seidenberg. The papers in the special issue were solicited from an outstanding group of connectionist language researchers, specifically to address the key subareas in connectionist language research and to discuss the future prospects of connectionist psycholinguistics. For the purpose of this book, each paper has been updated, including the addition of a descriptive list of further readings. In most cases the papers have also been substantially revised.

Part I, The State of the Art, brings us to the forefront of current connectionist modeling of psycholinguistic processing, with individual chapters on speech perception, morphology, sentential recursion, sentence processing, language production, and reading, beginning with an in-depth perspective on the breadth and variety of work in connectionist modeling of language. Part II, Future Prospects, provides a multifaceted discussion of the prospects for future research within connectionist psycholinguistics.

Each chapter is written by leading researchers who are defining the current state of the art within the connectionist approach to language. The book should therefore provide both a summary of where the field stands and a stimulus to future research in connectionist psycholinguistics. More generally, the book is aimed at researchers, scholars, and advanced students in psychology, linguistics, psycholinguistics, cognitive neuroscience, cognitive science, philosophy, or computer science with interest in the psychology of language and in computational approaches to the understanding of psycholinguistic processing.

1

Connectionist Psycholinguistics: The Very Idea

Morten H. Christiansen and Nick Chater

What is the significance of connectionist models of language processing? Will connectionism ultimately replace, complement, or simply implement the symbolic approach to language? Early connectionist models attempted to address this issue by showing that connectionist models could, in principle, capture aspects of language processing and linguistic structure. Little attention was generally paid to the modeling of data from psycholinguistic experiments. However, we suggest that connectionist language processing has matured and that the field is now moving forward into a new phase in which closer attention is paid to detailed psycholinguistic data. This book provides the first comprehensive overview of work within the emergent field of "connectionist psycholinguistics," connectionist models that make close contact with psycholinguistic results.

But how are we to assess the models within this emerging new area of research? We suggest that computational models of psycholinguistic processing, whether connectionist or symbolic, should attempt to fulfill three criteria: (1) data contact, (2) task veridicality, and (3) input representativeness (Christiansen & Chater, 2001). Data contact refers to the degree to which a model provides a fit with psycholinguistic data. We distinguish here between primary and secondary data contact. Primary data contact involves fitting results from specific psycholinguistic experiments (e.g., reaction-time data), whereas secondary data contact involves fitting general patterns of behavior (e.g., experimentally attested developmental changes in language processing) rather than specific results. Task veridicality refers to the degree of match between the task facing people and the task given to the model. Although a precise match is typically difficult to obtain, it is important to minimize the discrepancy. For example, much early work on modeling the English past tense suffers from low task veridicality (e.g., Rumelhart & McClelland, 1986, but see, e.g., Hoeffner, 1997, for an exception) because models are trained to map verb stems to past-tense forms, a task unlikely to be relevant to children's language acquisition. Input representativeness refers to the degree to which the information given to the model reflects what is available to a person or child. For example, the computational modeling of morphology suffers from the lack of good training corpora of high input representativeness with which to train the models. This problem is most serious for non-English morphology, making it problematic to make a priori conclusions about the feasibility of connectionist accounts in the area (e.g., Berent, Pinker, & Shimron, 1999).

It is also important to take stock of where symbolic models stand on our three criteria for computational psycholinguistics. Interestingly, few symbolic models make direct contact with psycholinguistic data. Most of the exceptions are within the study of sentence processing, where some comprehensive models of word-by-word reading times exist (e.g., Gibson, 1998; Just & Carpenter, 1992) and have a reasonable degree of task veridicality. More generally, however, symbolic models appear to pay little attention to task veridicality. Indeed, the rule-based models of the English past tense (e.g., Pinker, 1991) involve the same stem-to-past-tense mappings as the early connectionist models, and thus suffer from the same low task veridicality. Input representativeness is often ignored in symbolic models, in part because learning plays a minimal role in the performance of these models, and in part because symbolic models tend to be focused on more abstract fragments of language, rather than the more realistic language input that some connectionist models can handle. Low input representativeness may, for these reasons, actually inflate performance for many types of symbolic models, whereas the opposite tends to be true of connectionist models.

Currently, then, connectionism appears to provide a better framework for detailed psycholinguistic modeling than the symbolic approach. For many connectionists the advantages of this framework for doing computational psycholinguistics derive from a number of properties of the connectionist models.

Learning. Connectionist networks typically learn from experience, rather than being fully prespecified by a designer. By contrast, symbolic computational systems, including those concerned with language processing, are typically, but not always, fully specified by the designer.

Generalization. Few aspects of language are simple enough to be learnable by rote. The ability of networks to generalize to cases on which they have not been trained is thus a critical test for many connectionist models.

Representation. Because they are able to learn, the internal codes used by connectionist networks need not be fully specified by a designer, but are devised by the network so as to be appropriate for the task. Developing methods for understanding the codes that the network develops is an important strand of connectionist research. While internal codes may be learned, the inputs and outputs to a network generally use a code specified by the designer. These codes can be crucial in determining network performance. How these codes relate to standard symbolic representations of language in linguistics is a major point of contention.

Rules versus Exceptions. Many aspects of language can be described in terms of what have been termed "quasi-regularities," regularities that are usually true but admit some exceptions. According to the symbolic descriptions used by modern linguistics, these quasi-regularities may be captured in terms of a set of symbolic rules and sets of exceptions to those rules. Symbolic models often incorporate this distinction by having separate mechanisms that deal with rule-governed and exceptional cases. It has been argued that connectionist models provide a single mechanism that can pick up general rules while learning the exceptions to those rules. While this issue has been a major point of controversy surrounding connectionist models, it is important to note that attempting to provide single mechanisms for rules and exceptions is not essential to the connectionist approach; one or both separate mechanisms for rules and exceptions could themselves be modeled in connectionist terms (Coltheart, Curtis, Atkins, & Haller, 1993; Pinker, 1991; Pinker & Prince, 1988). A further question is whether networks really learn rules at all, or whether they simply approximate rulelike behavior. Opinions differ concerning whether the latter is an important positive proposal, which may lead to a revision of the role of rules in linguistics (Rumelhart & McClelland, 1986; see also Smolensky, 1988), or whether it is a fatal problem with connectionist models of language processing (Marcus, 1998; Pinker & Prince, 1988).

These four properties all play important roles in the models described in Part I of this volume, as well as in the appraisals of connectionist psycholinguistics presented in Part II.

ORGANIZATION OF THIS VOLUME

Part I of this volume, The State of the Art, presents the current state of the art in connectionist psycholinguistics with specific models from five key areas: speech processing, morphology, sentence processing, language production, and reading aloud. Part II, Future Prospects, then provides three contrasting perspectives on the field from leading researchers working on computational models of human natural language processing.

Part I begins with Chapter 2, Connectionist Psycholinguistics in Perspective, by Morten H. Christiansen and Nick Chater. This chapter provides an in-depth introduction to the field of connectionist psycholinguistics, and sets the context for the rest of the volume. The historical roots of the connectionist approach to language processing are traced and key themes that arise throughout different areas of connectionist psycholinguistics are highlighted. The chapter also provides a detailed review of the five key empirical areas described in the chapters comprising the rest of this part of the book, highlighting the interplay between connectionist modeling and empirical research. This review indicates that connectionist psycholinguistics has already had a significant impact on the psychology of language, and suggests that connectionist models are likely to have an important influence on future research. With this review in place, the subsequent chapters in this part of the book present central recent developments in the field.

Chapter 3, Simulating Parallel Activation in Spoken Word Recognition, by M. Gareth Gaskell and William D. Marslen-Wilson, concerns the connectionist modeling of speech processing. A critical property of the perception of spoken words is the transient ambiguity of the speech signal. Speech information is spread out across time, and early on in the processing of a word the speech information will be compatible with more than one lexical item. In localist models of speech perception this property is captured by allowing the parallel activation of multiple independent lexical representations. Gaskell and Marslen-Wilson examine how this property can be accommodated in a distributed model of speech perception, in which word representations are not independent. In this case an approximation to the activation of more than one representation is possible by activating a "blend" of the different distributed representations. Statistical analyses of vector spaces show that coactivation of multiple distributed representations is inherently noisy, and depends on parameters such as sparseness and dimensionality. Furthermore, the characteristics of coactivation vary considerably, depending on the organization of distributed representations within the mental lexicon. This view of lexical access is supported by analyses of phonological and semantic word representations, which provide an explanation of a recent set of experiments on coactivation in speech perception (Gaskell & Marslen-Wilson, 1999). More generally, this work illustrates a tight interplay between connectionist psycholinguistic modeling and experimental psycholinguistic research. Thus, the model provides for a good primary data contact and reasonable input representativeness, but suffers from a relatively poor task veridicality because of the abstract nature of the simulations.

Chapter 4, A Connectionist Model of English Past-Tense and Plural Morphology, by Kim Plunkett and Patrick Juola, concerns what has been one of the most controversial domains to which connectionist research has been applied: morphological processing. Theorists advocating a symbolic perspective have frequently taken morphology as a paradigmatic case of a "rule + exception" mapping. A rigid symbolic rule, which specifies a regular morphological mapping, is presumed to be supplemented with a set of explicit exceptions, which are assumed to be processed by a very different mechanism. In line with much connectionist work in this area, Plunkett and Juola make the opposite assumption, that a single mechanism explains both rule and exception cases in morphological processing. Specifically, they model the acquisition of English noun and verb morphology using a single connectionist network. The network is trained to produce the plurals and past-tense forms for a large corpus of monosyllabic English nouns and verbs. The developmental trajectory of network performance is analyzed in detail and is shown to mimic a number of important features of the acquisition of English noun and verb morphology in young children. These include an initial error-free period of performance on both nouns and verbs followed by a period of intermittent overregularization of irregular nouns and verbs. Errors in the model show evidence of phonological conditioning and frequency effects. Furthermore, the network demonstrates a strong tendency to regularize denominal verbs and deverbal nouns and masters the principles of voicing assimilation. Despite being dealt with by a single network, nouns and verbs exhibit some important differences in their profiles of acquisition. Most important, noun inflections are acquired earlier than verb inflections. The simulations generate several empirical predictions that can be used to further evaluate the suitability of this type of cognitive architecture in the domain of inflectional morphology, thus pointing the way for close links between computational and empirical research. The model has good secondary data contact and decent input representativeness, but the task veridicality is poor because the task of mapping noun and verb stems to plural and past-tense inflections is not likely to play a large role in language acquisition.

Chapter 5, Finite Models of Infinite Language: A Connectionist Approach to Recursion, by Morten H. Christiansen and Nick Chater, deals with another theoretical issue that has been seen as strongly supporting a symbolic, rather than a connectionist, approach to language processing: natural language recursion. Since the inception of modern linguistics there has been considerable emphasis on the importance of recursive phenomena in natural language, and the assumption that any approach to sentence processing must allow for unbounded recursion. Indeed, the existence of different kinds of recursion has had important implications on the choices of symbolic formalisms (e.g., different kinds of generative grammars, different classes of parser and generator) that have been used to explain natural language. From this perspective, natural language recursion presents a difficult challenge to any nonsymbolic account of natural language processing.

A range of connectionist approaches have been put forward that attempt to deal with recursion in natural language, although they have not typically achieved the unbounded character of natural language recursion that linguists typically assume. Christiansen and Chater note, though, that the proposition that natural language allows unbounded applications of recursion may make an inappropriate target for connectionist modeling. Instead, they argue that a more appropriate goal for connectionism is to account for the levels of performance that people exhibit when exposed to recursive constructions—to address recursion as a purely psycholinguistic phenomenon, rather than as a linguistic abstraction. It is important to note that people's ability to process recursive constructions is quite limited. People produce only a very limited number of complex recursive constructions in naturally occurring speech, and this is reflected in the empirically documented difficulties that people experience when processing such structures.

Christiansen and Chater present a connectionist model of human performance in processing recursive language structures, based on Elman's (1990) simple recurrent network (SRN). The model is trained on simple artificial languages inspired by Chomsky (1957). They find that the qualitative performance profile of the model closely matches human behavior, both on the relative difficulty of center-embedded and cross-dependency, and between the processing of these complex recursive structures and right-branching recursive constructions. Christiansen and Chater analyze how these differences in performance are reflected in the internal representations of the model by performing discriminant analyses on these representations both before and after training. The model has good primary data contact and reasonable task veridicality, but the input representativeness is low because of the abstractness of the artificial languages. More generally, this work suggests a novel explanation of people's limited recursive performance, without assuming the existence of a mentally represented grammar allowing unbounded recursion.

Chapter 6, Dynamical Systems for Sentence Processing, by Whitney Tabor and Michael K. Tanenhaus, like the previous chapter, addresses the question of natural language processing at the level of the sentence, using input patterned on natural language rather than the more abstract structures used by Christiansen and Chater. Tabor and Tanenhaus suggest that the theory of dynamical systems, originated in the physical sciences, provides a revealing general framework for modeling the representations and mechanisms underlying sentence processing. Recent work in sentence processing (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998) suggests that parsing biases change fairly continuously over the course of processing the successive words of a sentence. Connectionist networks are good at fitting graded data, and their dynamical properties are naturally suited to modeling continuously changing quantities. But the connectionist network that has been most successful in modeling natural language syntax (Elman's SRN, which is used by Christiansen and Chater in the previous chapter) does not explicitly model processing times. They argue that, like many connectionist models at the present time, the SRN is analytically opaque: It is difficult to see the principles underlying its solutions to complex tasks. And it is relativistic—no categorical distinctions are made between grammatical and ungrammatical strings—so it is hard to use linguistic structural insights, which make heavy use of such distinctions, to get past the opaqueness. They suggest that dynamical systems theory, through its insight into the relationship between quantitative and topological properties of systems, offers a solution to these shortcomings.

As in their previous work (Tabor, Juliano, & Tanenhaus, 1997), Tabor and Tanenhaus add a postprocessor to the SRN that has explicit dynamics, thus introducing potentially useful dynamical systems concepts: attractors, basins, saddle points, trajectories. They call this the Visitation Set Gravitation (VSG) model. Trained on a simple formal language that shares certain key properties with English, the model predicts the important reading-time contrasts in a recent study of the real-time evolution of parsing biases (McRae et al., 1998).

Further examination of the VSG model reveals that a standard structural contrast in dynamical systems—between saddle points and attractors—maps onto the fundamental linguistic contrast between ungrammatical and grammatical strings, thus helping to bridge the gap between connectionist models and linguistic theory. And without further modification of the model, a behaviorally accurate analysis of semantically strange sentences falls out: They are grammatical sentences that involve long trajectories in the state space of the dynamical system and thus have long processing times. This insight helps move work on formal language learning in the much-needed direction of addressing semantic structure.

Overall, the Tabor and Tanenhaus model has good primary data contact and decent task veridicality, but the input representativeness is low because of the simplicity of their formal language. The results suggest that dynamical-systems theory is a promising source of ideas for relating the flexible, real-time behavior of the human language processor to its overarching, relatively static, categorical organization. This application of dynamical-systems ideas is part of a larger movement within cognitive science (e.g., Kelso, 1997; Port & van Gelder, 1995; Thelen & Smith, 1994), which seeks to understand cognition in dynamical terms. Language processing provides a challenging test case for the application of the dynamical approach, because language has traditionally been conceived from a symbolic perspective. It is interesting, too, to wonder to what extent connectionist researchers will follow Tabor and Tanenhaus in using ideas from dynamical-systems theory to construct and understand connectionist systems. If this does occur, it might represent a substantial departure from the current technical literature on connectionist networks, which is grounded in probability, information theory, and statistical mechanics, rather than dynamical ideas (Bishop, 1995; Frey, 1998).

Chapter 7, Connectionist Models of Language Production: Lexical Access and Grammatical Encoding, by Gary S. Dell, Franklin Chang, and Zenzi M. Griffin, moves the focus from how language is understood to how it is produced. Language production, like language understanding, has frequently been characterized as involving the operation of symbolic processes on a generative grammar, and a specification of the message to be produced, in terms of a symbolically encoded underlying "logical form" or "conceptual representation." In contrast to this kind of account, there has also been a long tradition of connectionist theorizing about language production. Indeed, Dell's (1986) "spreading activation" model of speech production was one of the most important models in the revival of interest in connectionist models of psychological processes, which began in the early to mid-1980s. In their chapter, Dell et al. describe the most recent developments in this approach to modeling speech production. Specifically, they outline the issues and challenges that must be addressed by connectionist models of lexical access and grammatical encoding, and review three recent models. The models illustrate the value of a spreading activation approach to lexical access in production, the need for sequential output in both phonological and grammatical encoding, and the potential for accounting for structural effects on phonological errors and syntactic priming. These models account for a broad range of data on speech production, from the analysis of speech errors, to the performance of aphasic patients, to results from syntactic priming studies. Indeed, in speech-production research the interplay between connectionist modeling and the gathering of empirical data that we view as constitutive of connectionist psycholinguistics is particularly well developed.

Dell et al. consider several specific models, rather than attempting a single overarching model of speech production. Individually, the models have good primary or secondary data contact and good task veridicality, but all models suffer from relatively low input representativeness because the models only cover small language fragments. An important question for future research concerns the degree to which models of specific aspects of speech production can be integrated in a cohesive way, an issue that also arises in relation to the models of speech and language processing described in earlier chapters of this book.

The chapters described so far have focused on the comprehension and production of speech, rather than how written language is processed. The reading of single words has, in particular, been an area of intense connectionist research. Chapter 8, A Connectionist Approach to Word Reading and Acquired Dyslexia: Extension to Sequential Processing, by David C. Plaut, outlines a new model of reading, building on the long research tradition. Plaut begins by discussing some general principles of the connectionist approach to word reading—of which he is a leading proponent—including distributed representation, graded learning of statistical structure, and interactivity in processing. These principles have led to the development of explicit computational models that account for an impressively broad range of data, from the interaction of word frequency and spelling-sound consistency in normal skilled reading to analogous effects in the reading errors of surface dyslexic patients and the co-occurrence of visual and semantic errors in deep dyslexia.

Plaut notes, though, that there have been recent empirical challenges to these models, and the approach in general, relating to the influence of orthographic length on the naming latencies of both normal and dyslexic readers. For instance, the models account for relatively little variance associated with individual words in large databases of naming latencies, partly due to insufficient sensitivity to orthographic length. The models also underestimate length effects in the naming latencies for nonwords. This kind of empirical challenge is an illustration of the productive interaction between connectionist modeling and empirical research—predictions of connectionist models have had a crucial impact in directing the search for relevant empirical confirmation or disconfirmation.

Plaut addresses this challenge by presenting a new model that generates sequential phonological output in response to written input. He trains an SRN (Elman, 1990) to produce sequences of single phonemes as output when given position-specific letters as input. The model was also trained to maintain a representation of its current position within the input string. When the model found a peripheral portion of the input difficult to pronounce, it used the position signal to refixate the input, shifting the peripheral portion to the point of fixation where the model has had more experience in generating pronunciations. In this way the model could apply the knowledge tied to the units at the point of fixation to any difficult portion of the input. Early on in training, the model required multiple fixations to read words, but as the model became more competent it eventually read most words in a single fixation. The model could also read nonwords about as well as skilled readers, occasionally falling back on a refixation strategy for difficult nonwords. The model exhibits an effect of orthographic length and a frequency-by-consistency interaction in its naming latencies. When subject to peripheral damage, the model exhibits an increased length effect that interacts with word frequency, characteristic of letter-by-letter reading in pure alexia. The model thus has a good primary data contact and good task veridicality, but input representativeness suffers somewhat because the model is only trained on monosyllabic words. Plaut notes that the model is not intended as a fully adequate account of all the relevant empirical phenomena. But the model provides a compelling demonstration of how connectionist models may be extended to provide deeper insight into sequential processes in reading.

Plaut's chapter concludes the first part of the book, which reviews current models of connectionist language processing. The second part of the book consists of three insightful perspectives on the significance, interpretation, and utility of connectionist psycholinguistics by eminent researchers in the cognitive science of language processing.

Chapter 9, Constraint Satisfaction in Language Acquisition and Processing, by Mark S. Seidenberg and Maryellen C. MacDonald, sets out the most radical connectionist agenda, seeing connectionism as potentially undermining classical symbolic theorizing in linguistics and psycholinguistics. In particular, they see connectionist psycholinguistics as part of a larger theoretical framework focusing on probabilistic constraints on language processing and language acquisition (Seidenberg, 1997). This probabilistic framework offers an alternative viewpoint on language and language use to that found in generative linguistics. The generative approach attempts to characterize knowledge of language (i.e., competence grammar) and then asks how this knowledge is acquired and used. Seidenberg and MacDonald's probabilistic approach is performance oriented: The goal is to explain how people comprehend and produce utterances and how children acquire this skill. From a probabilistic perspective, using language is thought to involve exploiting multiple probabilistic constraints over various types of linguistic and nonlinguistic information. Children begin accumulating this information at a young age. The same constraint-satisfaction processes that are central to language use in adulthood also serve as the bootstrapping processes that provide entry into language in childhood. Framing questions about acquisition in terms of models of skilled language use has important consequences for arguments concerning language learnability and holds out the possibility of a unified theory of acquisition and use.

Seidenberg and MacDonald put forward a vigorous case for opposition between connectionist and symbolic approaches to language. But this is, of course, by no means the only possible viewpoint. Language might instead be viewed as emerging from a mixture of linguistic rules, which can be specified in symbolic terms, and probabilistic factors that determine how language is used in specific contexts; and, indeed, symbolic linguistic rules need not, perhaps, be quite as rigid as is typically assumed. Thus, a more conciliatory line between connectionist psycholinguistics and symbolic, generative linguistics may be imagined.

Chapter 10, Grammar-Based Connectionist Approaches to Language, by Paul Smolensky, outlines a specific conception of how connectionist and symbolic theorizing about language might be integrated, rather than set against each other. In particular, Smolensky argues that connectionist research on language can and must involve the development of grammar formalisms rather than merely producing connectionist computer models. From formulations of the fundamental theoretical commitments of connectionism and of generative grammar, it is argued that these two paradigms are mutually compatible: The commitments of connectionism concern computational principles, and those of generative grammar concern explanations of certain fundamental empirical characterizations of human language. Integrating the basic assumptions of the two paradigms results in formal theories of grammar that centrally incorporate a certain degree of connectionist computation. Two such grammar formalisms—Harmonic Grammar (Legendre, Miyata, & Smolensky, 1990) and Optimality Theory (Prince & Smolensky, 1997)—are briefly introduced to illustrate grammar-based approaches to connectionist language research. The strengths and weaknesses of grammar-based research and more traditional model-based research are argued to be complementary: Grammar-based research more readily allows explanations of general patterns of language, while model-based research more readily allows exploration of the full richness and power of connectionist computational mechanisms. This complementarity of strengths suggests a significant role for both strategies in the spectrum of connectionist language research.

Smolensky's standpoint provides a counterweight to the view that connectionist psycholinguistics should attempt to overturn previous theorizing about language and language processing. Moreover, the synthesis of principles from connectionism and generative grammar that he outlines gains considerable credibility from its very widespread influence in modern linguistics. Indeed, optimality theory is widely viewed within linguistics as one of the central theoretical developments within the field. Smolensky's approach may not, however, satisfy more radical connectionists, who may see the move from specific and implementable connectionist models of psychological processes (such as are described in the first half of this book) to abstract connectionist principles as too great a departure from the original aims of the connectionist paradigm.

Chapter 11, Connectionist Sentence Processing in Perspective, by Mark Steedman, presents an outsider's perspective on the project of connectionist psycholinguistics. Steedman has been associated with symbolic approaches to language, and has been involved in pioneering novel linguistic formalisms, such as categorial grammar, as well as carrying out highly influential computational and experimental work on human language processing. Steedman focuses on connectionist sentence processing, a topic discussed in Chapters 5 and 6 of this book. Steedman argues that the emphasis in the connectionist sentence-processing literature on distributed representation and emergence of grammar from such systems seems to have prevented connectionists and symbolic theorists alike from recognizing the often close relations between their respective systems. He argues that SRN models (Elman, 1990) are more directly related to stochastic Part-of-Speech taggers than to parsers or grammars as such, while recursive auto-associative memory of the kind pioneered by Pollack (1990) and incorporated in many hybrid connectionist parsers since may be useful for grammar induction from a network-based conceptual structure as well as for structure building.

These observations suggest some interesting new directions for connectionist sentence-processing research, including more efficient devices for representing finite-state machines, and acquisition devices based on a distinctively connectionist-grounded conceptual structure. Thus, Steedman, like Smolensky, argues for an integration of connectionist and symbolic views of language and language processing. But Steedman and Smolensky differ concerning the nature of the integration. Whereas Smolensky argues that connectionist principles should be integrated into grammar formalisms, Steedman sees connectionist networks as integrating with symbolic language-processing mechanisms to produce a hybrid computational account of language processing and acquisition. And both Smolensky and Steedman differ from the more radical agenda of Seidenberg and MacDonald, which aims to replace, rather than interconnect with, previous theories of language processing and structure. Clearly, only the future development of connectionist research will decide which of these perspectives, each of which is persuasively argued, proves to be the most fruitful.

THE SIGNIFICANCE OF CONNECTIONIST PSYCHOLINGUISTICS

Current connectionist models involve important simplifications with respect to real natural language processing. In some cases these simplifications are relatively modest. For example, models of reading aloud typically ignore how eye movements are planned and how information is integrated across eye movements; they also tend to ignore the sequential character of speech output, and typically deal only with short words. In other cases the simplifications are more drastic. For example, connectionist models of syntactic processing involve vocabularies and grammars that are vastly simplified. However, it is important to note that symbolic models in many cases have lower task veridicality and input representativeness than their connectionist counterparts. Furthermore, many symbolic models may give the appearance of good data contact simply because they have not yet been implemented and have therefore not been tested in an empirically rigorous way, in contrast to the connectionist models.

The present breadth and significance of connectionist psycholinguistics, as evidenced by the chapters in this volume, suggests that the approach has considerable potential. Despite some attempts to argue for a priori limitations on connectionist language processing (e.g., Pinker & Prince, 1988; Marcus, 1998), connectionist psycholinguistics has already had a major impact on the psychology of language.

First, connectionist models have provided the first fully explicit and psychologically relevant computational models in a number of language-processing domains, such as reading and past-tense learning. Previous accounts in these areas consisted of "box-and-arrow" flow diagrams rather than detailed computational mechanisms. Whatever the lasting value of connectionist models themselves, they have certainly raised the level of theoretical debate in these areas by challenging theorists of all viewpoints to provide computationally explicit accounts.

Second, the centrality of learning in connectionist models has brought a renewed interest in mechanisms of language learning (Bates & Elman, 1993). While Chomsky (e.g., 1986) has argued that there are "universal" aspects of language that are innate, the vast amount of information specific to the language that the child acquires must be learned. Connectionist models provide mechanisms for how (at least some of) this learning might occur, whereas previous symbolic accounts of language processing have not taken account of how learning might occur. Furthermore, the attempt to use connectionist models to learn syntactic structure encroaches on the area of language for which Chomsky has argued innate information must be central. The successes and failures of this program thus directly bear on the validity of this viewpoint.

Third, the dependence of connectionist models on statistical properties of their input has been one contributing factor in the upsurge of interest in the role of statistical factors in language learning and processing (MacWhinney, Leinbach, Taraban, & McDonald, 1989; Redington & Chater, 1998). This renewed interest in the statistical properties of language and statistical methods of analysis is, of course, entirely compatible with the view that language processing takes account of structural properties of language, as described by classical linguistics. But more radical connectionists have, as we have noted, also attempted to encroach on the territory of classical linguistics.

Finally, connectionist systems have given rise to renewed theoretical debate concerning what it really means for a computational mechanism to implement a rule, whether there is a distinction between "implicit" and "explicit" rules (see, e.g., Davies, 1995, for discussion), and which kind should be ascribed to the human language-processing system.

The potential implications of a realistic connectionist approach to language processing are enormous. If realistic connectionist models of language processing can be provided, then the possibility of a radical rethinking not just of the nature of language processing, but of the structure of language itself, may be required. It might be that the ultimate description of language resides in the structure of complex networks, and it can only be approximately expressed in terms of structural rules, in the style of generative grammar (Seidenberg & MacDonald, Chapter 9, this volume). On the other hand, it may be that connectionist models can only succeed to the extent that they build in standard linguistic constructs (Smolensky, Chapter 10, this volume), or fuse with symbolic models to create a hybrid approach (Steedman, Chapter 11, this volume). We suggest that the only way to determine the ultimate value of connectionist psycholinguistics is to pursue it with the greatest possible creativity and vigor, as exemplified by the chapters in this volume.

NOTE

We are grateful to the Cognitive Science Society for allowing us to put this book together based on papers that appeared in *Cognitive Science*. All the chapters except one are revised and updated versions of earlier solicited papers found in the special issue, "Connectionist models of human language processing: Progress and prospects," vol. 23, no. 4, 1999, edited by Morten Christiansen, Nick Chater, and Mark Seidenberg. Chapter 5 is a substantially revised version of an article that appeared in vol. 23, no. 2, 1999. We would like to thank the contributors for updating and revising their papers for this book, and to the reviewers for helping to ensure a very high quality of papers throughout.

This work was partially supported by the Leverhulme Trust and by European Commission grant RTN-HPRN-CT-1999-00065 to Nick Chater.

REFERENCES

Bates, E. A., & Elman, J. L. (1993). Connectionism and the study of change. In M. J. Johnson (Ed.), *Brain development and cognition* (pp. 623–642). Cambridge, MA: Basil Blackwell.

Berent, I., Pinker, S., & Shimron, J. (1999). Default nominal inflection in Hebrew: Evidence for mental variables. *Cognition, 72*, 1–44.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York: Praeger.

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences, 5*, 82–88.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100*, 589–608.

Davies, M. (1995). Two notions of implicit rules. In J. E. Tomberlin (Ed.), *Philosophical perspectives: Vol. 9. AI, connectionism, and philosophical psychology*. Atascadero, CA: Ridgeview.

Dell, G. S. (1986). A spreading activation theory of retrieval in language production. *Psychological Review, 93*, 283–321.

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

Frey, B. J. (1998). *Graphical models for machine learning and digital communication*. Cambridge, MA: MIT Press.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). *Representation and competition in the processing of spoken words*. Manuscript submitted for publication.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*, 1–76.

Hoeffner, J. (1997). *Are rules a thing of the past? A single mechanism account of English past tense acquisition and processing*. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 98*, 122–149.

Kelso, J.A.S. (1997). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Harmonic grammar—a formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 388–395). Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language, 28*, 255–277.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology, 37*, 243–282.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*, 282–312.

Pinker, S. (1991). Rules of language. *Science, 253*, 530–535.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73–193.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence, 46*, 77–105.

Port, R. F., & van Gelder, T. (Eds.). (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science, 275*, 1604–1610.

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

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275*, 1599–1603.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1–74.

Tabor, W., Juliano, C., & Tanenhaus, M. K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes, 12*, 211–271.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

This page intentionally left blank

PART I

THE STATE OF THE ART

This page intentionally left blank

2

Connectionist Psycholinguistics in Perspective

Morten H. Christiansen and Nick Chater

Connectionist approaches to language have been, and still are, highly controversial. Some have argued that natural language processing from phonology to semantics can be understood in connectionist terms; others have argued that no aspects of natural language can be captured by connectionist methods. And the controversy is particularly heated because of the revisionist claims of some connectionists: For many, connectionism is not just an additional method for studying language processing; it also offers an alternative to traditional theories, which describe language and language processing in symbolic terms. Indeed, Rumelhart and McClelland (1987, p. 196) suggest "that implicit knowledge of language may be stored among simple processing units organized into networks. While the behavior of such networks may be describable (at least approximately) as conforming to some system of rules, we suggest that an account of the fine structure of the phenomena of language and language acquisition can best be formulated in models that make reference to the characteristics of the underlying networks." We shall see that the degree to which connectionism supplants, rather than complements, existing approaches to language is itself a matter of debate. Finally, the controversy over connectionist approaches to language is an important test case for the validity of connectionist methods in other areas of psychology.

In this chapter we aim to set the scene for the present volume on connectionist psycholinguistics, providing a brief historical and theoretical background as well as an update on current research in the specific topic areas outlined later. First we describe the historical and intellectual roots of connectionism, then introduce the elements of modern connectionism and how it has been applied to natural language processing, and outline some of the theoretical claims that have been made for and against it. We then consider five central topics within connectionist psycholinguistics: speech processing, morphology, sentence processing, language production, and reading. We evaluate the research in each of these areas in terms of the three criteria for connectionist psycholinguistics discussed in Chapter 1: data contact, task veridicality, and input representativeness. The five topics illustrate the range of connectionist research on language discussed in more depth in the other chapters in Part I of this volume. They also provide an opportunity to assess the strengths and weaknesses of connectionist methods across this range, setting the stage for the general debate concerning the validity of connectionist methods in Part II of this volume. Finally, we sum up and consider the prospects for future connectionist research, and its relation to other approaches to the understanding of language processing and linguistic structure.

BACKGROUND

From the perspective of modern cognitive science, we tend to see theories of human information processing as borrowing from theories of information processing in machines (i.e., from computer science). Within computer science, symbolic processing on general-purpose digital computers has proved to be the most successful method of designing practical computational devices. It is therefore not surprising that cognitive science, including the study of language processing, has aimed to model the mind as a symbol processor.

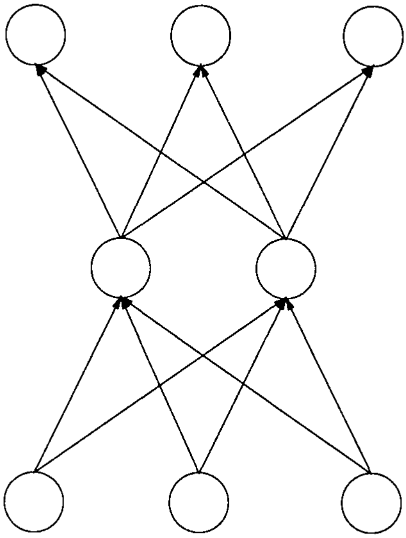
Historically, however, theories of human thought inspired attempts to build computational devices, rather than the other way around. Mainstream computer science arises from the intellectual tradition of viewing human thought as a matter of symbol processing. This tradition can be traced to Boole's (1854) suggestion that logic and probability theory describe "Laws of Thought," and that reasoning in accordance with these laws can be conducted by following symbolic rules. It runs through Turing's (1936) argument that all human thought can be modeled by symbolic operations on a tape (the Turing machine), through von Neumann's motivation for the design for the modern digital computer, to the development of symbolic computer programming languages, and thence to modern computer science, artificial intelligence, and symbolic cognitive science.

Connectionism (also known as "parallel distributed processing," "neural networks," or "neurocomputing") can be traced to a different tradition, which attempts to design computers inspired by the structure of the brain.[1] McCulloch and Pitts (1943) provided an early and influential idealization of neural function. In the 1950s and 1960s Ashby (1952), Minsky (1954), Rosenblatt (1962), and others designed computational schemes based on related idealizations. Aside from their biological origin, these schemes were of interest because they were able to learn from experience, rather than being designed. Such "self-organizing" or learning machines therefore seemed plausible as models of learned cognitive abilities, including many aspects of language processing (although Chomsky, 1965, among others, challenged the extent to which language is learned). Throughout this period connectionist and symbolic computation stood as alternative paradigms for modeling intelligence, and it was unclear which would prove to be the most successful. But gradually the symbolic paradigm gained ground, providing powerful models in core domains such as language (Chomsky, 1965) and problem solving (Newell & Simon, 1972). Connectionism was largely abandoned, particularly in view of the limited power of then current connectionist methods (Minsky & Papert, 1969). But more recently, some of these limitations have been overcome (e.g., Hinton & Sejnowski, 1986; Rumelhart, Hinton, & Williams, 1986), reopening the possibility that connectionism constitutes an alternative to the symbolic model of thought.

So connectionism is inspired by the structure and processing of the brain. What does this mean in practice? At a coarse level of analysis, the brain can be viewed as consisting of a very large number of simple processors, neurons, which are densely interconnected into a complex network. These neurons do not appear to tackle information processing problems alone. Rather, large numbers of neurons operate cooperatively and simultaneously to process information. Furthermore, neurons appear to communicate numerical values (encoded by firing rate), rather than passing symbolic messages, and, to a first approximation at least, neurons can be viewed as mapping a set of numerical inputs (delivered from other neurons) onto a numerical output (which is then transmitted to other neurons). Connectionist models are designed to mimic these properties: Hence, they consist of large numbers of simple processors, known as *units* (or nodes), which are densely interconnected into a complex network, and which operate simultaneously and cooperatively to solve information-processing problems. In line with the assumption that real neurons are numerical processors, units are assumed to pass only numerical values rather than symbolic messages, and the output of a unit is usually assumed to be a numerical function of its inputs.

The most popular of the connectionist networks is the *feed-forward network*, as illustrated in Figure 2.1. In this type of network the units are divided into "layers" and activation flows in one direction through the network, starting at the layer of input units and finishing at the layer of output units. The internal layers of the network are known as hidden units (HU). The activation of each unit is determined by its current input (calculated as

Figure 2.1
Feed-Forward Network



Information flows entirely bottom-up in these networks, from the input units through the hidden units to the output units, as indicated by the arrows.

the weighted sum of its inputs, as before). Specifically, this input is "squashed," so that the activation of each unit lies between 0 and 1. As the input to a unit tends to positive infinity, the level of activation approaches 1; as the input tends to negative infinity, the level of activation approaches 0. With occasional minor variations, this description applies equally to almost all feed-forward connectionist networks.

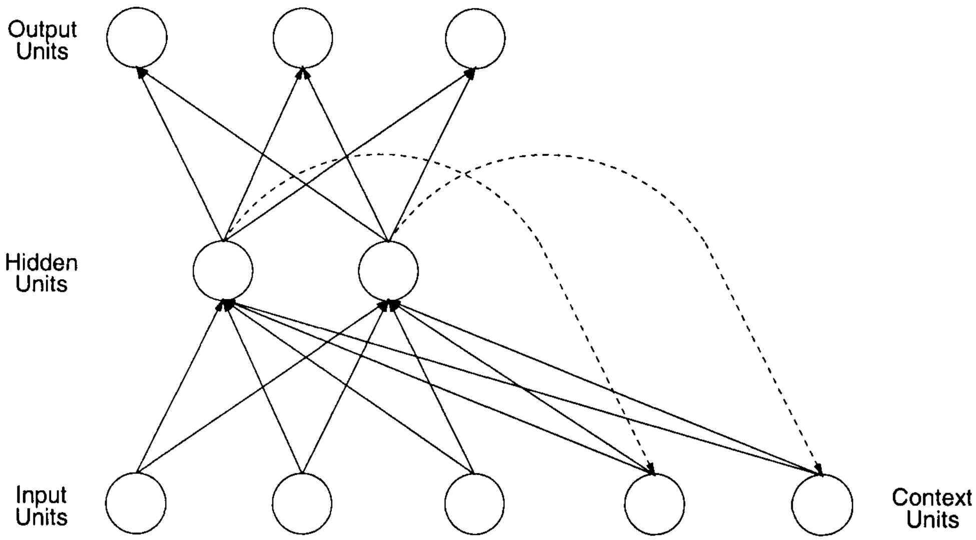
Feed-forward networks learn from exposure to examples, and learning is typically achieved using the back-propagation learning algorithm (Rumelhart et al., 1986; prefigured in Bryson & Ho, 1975; Werbos, 1974). When each input is presented, it is fed through the network and the output is derived. The output is compared against the correct "target" value and the difference between the two is calculated for each output unit. The squared differences are summed over all the output units to give an overall measure of the "error" that the network has made. The goal of learning is to reduce the overall level of error, averaged across input–target pairs. Back-propagation is a procedure that specifies how the weights of the network (i.e., the strengths of the connections between the units) should be adjusted in order to decrease the error. Training with back-propagation is guaranteed (within certain limits) to reduce the error made by the network. If everything works well, then the final level of error may be small, meaning that the network produces the desired output. Notice that the network will produce an output not only for inputs on which it has been trained, but for any input. If the network has learned about regularities in the mapping between inputs and targets, then it should be able to generalize successfully (i.e., to produce appropriate outputs in response to these new inputs).

Back-propagation may sound too good to be true. But note that back-propagation merely guarantees to adjust the weights of the network to reduce the error; it does not guarantee to reduce the error to 0, or a value anywhere near 0. Indeed, in practice, back-propagation can configure the network so that error is very high, but changes in weights in any direction lead to the same or a higher error level, even though a quite different configuration of weights would give rise to much lower error, if only it could be found by the learning process. The network is stuck in a *local minimum* in weight space, and cannot find its way to better local minima, or better still, to the optimal weights that are the global minimum for error. Attempting to mitigate the problem of local minima is a major day-to-day concern of connectionist researchers, as well as being a focus of theoretical research. The problem of local minima can be reduced by judicious choice among the large number of variants of back-propagation, and by appropriate decisions on the numerous parameters involved in model building (such as the number of hidden units used, whether learning proceeds in small or large steps, and many more). But the adjustment of these parameters is often more a matter of judgment, experience, and guesswork than it is a product of theoretical analysis. Despite these problems, back-propagation is surprisingly successful in many contexts. Indeed, the feasibility of back-propagation learning has been one of the reasons for the renewed interest in connectionist research. Prior to the discovery of back-propagation, there were no well-justified methods for training multilayered networks. The restriction to single-layered networks was unattractive, since Minsky and Papert (1969) showed that such networks, sometimes known as "perceptrons," have very limited computational power. It is partly for this reason that hidden units are viewed as having such central importance in many connectionist models; without hidden units, most interesting connectionist computation would not be possible.

A popular variation of the feed-forward network is the simple recurrent network (SRN; Elman, 1988, 1990) (see Figure 2.2). This network is essentially a standard feed-forward network equipped with an extra layer of so-called context units. At a particular time step an input pattern is propagated through the hidden-unit layer to the output layer (solid arrows). At the next time step the activation of the hidden-unit layer at the previous time step is copied back to the context layer (dashed arrows) and paired with the current input (solid arrows).[2] This means that the current state of the hidden units can influence the processing of subsequent inputs, providing a limited ability to deal with integrated sequences of input presented successively.

Whereas simple recurrent networks can be trained using the standard back-propagation learning algorithm, fully recurrent networks are trained using more complex learning algorithms, such as discrete back-propagation

Figure 2.2
Simple Recurrent Network



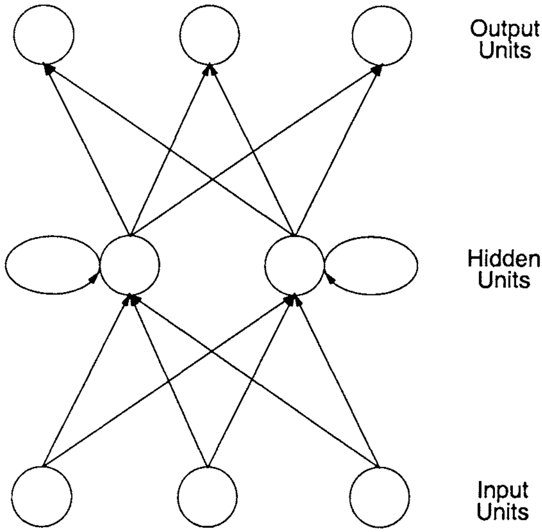
At a particular time step an input pattern is propagated through the hidden-unit layer to the output layer (solid arrows). At the next time step the activation of the hidden-unit layer at the previous time step is copied back to the context layer (dashed arrows) and paired with the current input (solid arrows).

through time (Williams & Peng, 1990) and continuous back-propagation (Pearlmutter, 1989). This type of network architecture is shown in Figure 2.3. Through the recurrent links (circular arrows), current activation can affect future activations similarly to the simple recurrent network, but in a more fine-grained manner and potentially reaching further back in time.

Another popular network architecture is the interactive activation network, shown in Figure 2.4. This type of network is completely prespecified (i.e., it does not learn). It consists of a sequence of unit layers. Units in the first layer typically encode fine-grained features of the input (e.g., visual or phonetic features). Units in the subsequent layers encode elements of increasingly higher levels of analyses (e.g., letters → words or phonemes → words). Units are connected using bidirectional links that can be either excitatory (arrows) or inhibitory (filled circles). This style of connectivity allows for activation to flow both bottom-up and top-down, reinforcing mutually consistent states of affairs and inhibiting mutually inconsistent states of affair.

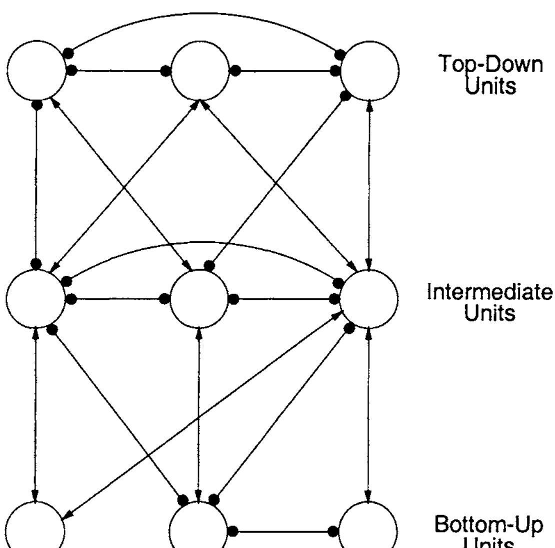
The behavior of individual units in interactive activation networks is somewhat more complex than in the network architectures we have described so far, because it depends not only on the current input but also on the previous

Figure 2.3
Fully Recurrent Network



Recurrent links (circular arrows) allow activation at the current time step to affect activations for many future time steps.

Figure 2.4
Interactive Activation Network



The links are bidirectional and can be either excitatory (arrows) or inhibitory (filled circles). Activation in this network flows both bottom-up and top-down.

level of activity of the unit. If the input to a unit is 0, then all that happens is that the level of activity of the unit decays exponentially. The input to the unit is, as is standard, simply the weighted sum of the outputs of the units that feed into that unit (where the weights correspond to the strengths of the connections). If the input is positive, then the level of activity is increased in proportion both to that input and to the distance between the current level of activation and the maximum activation (conventionally set at 1); if the input is negative, the level of activity is decreased in proportion to the input and to the distance between the current level of activation and the minimum activation (conventionally set at -1).

While this behavior sounds rather complex, the basic idea is simple. Given a constant input, the unit will gradually adjust to a stable level where the exponential decay balances with the boost from that input: Positive constant inputs will be associated with positive stable activation, negative constant inputs with negative stable activation; small inputs lead to activations levels close to 0, while large inputs lead to activation values which tend to be 1 or -1. If we think of a unit as a feature detector, then an activation level near 1 corresponds to a high level of confidence that the feature is present; an activation level near -1 corresponds to a high level of confidence that it is not.

With respect to the relationship between connectionist models and the brain, it is important to note that none of the connectionist architectures that we have described amount to realistic models of brain function (see, e.g., Sejnowski, 1986). They are unrealistic at the level of individual processing units, where the models not only drastically oversimplify, but knowingly falsify, many aspects of the function of real neurons, and in terms of the structure of the connectionist networks, which bear little if any relation to brain architecture. One avenue of research is to seek increasing biological realism (e.g., Koch & Segev, 1989). In the study of the areas of cognition in which few biological constraints are available, most notably language, researchers have concentrated on developing connectionist models with the goal of accurately modeling human behavior. They therefore take their data from cognitive psychology, linguistics, and cognitive neuropsychology, rather than from neuroscience. Thus, connectionist research on language appears to stand in direct competition with symbolic models of language processing.

As noted earlier, the relative merits of connectionist and symbolic models of language are hotly debated. But should they be viewed as in opposition at all? After all, advocates of symbolic models of language processing assume that symbolic processes are somehow implemented in the brain. Thus, they too are connectionists at the level of implementation. But symbolic theorists assume that language processing can be described at two levels: at the psychological level, in terms of symbol processing, and at the implementational level, in neuroscientific terms (to which connectionism is a crude approximation). If this is right, then connectionist modeling should proceed by taking symbol-processing models of language processing and attempting to implement these in connectionist networks. Advocates of this view (Fodor & Pylyshyn, 1988; Marcus, 1998; Pinker & Prince, 1988) typically assume that it implies that symbolic modeling should be entirely autonomous from connectionism; symbolic theories set the goalposts for connectionism, but not the other way round. Chater and Oaksford (1990) have argued that even according to this view there will be two-way influences between symbolic and connectionist theories, since many symbolic accounts can be ruled out precisely because they could not be neurally implemented. But most connectionists in the field of language processing have a more radical agenda: not to implement, but to challenge, to varying degrees, the symbolic approach to language processing. Part II of this book will illustrate a variety of contemporary viewpoints on the relationship between connectionist and symbolic theories of language.

With these general issues in mind, let us now consider the broad spectrum of connectionist models of language processing.

SPEECH PROCESSING

Speech processing in its broadest sense encompasses a broad range of cognitive processes, from those involved in low-level acoustical analysis to those involved in semantic and pragmatic interpretation of utterances. Here we shall focus much more narrowly, on the processes involved in segmenting and recognizing spoken words from input that is represented in a linguistic form (e.g., as sequences of phonetic features or phonemes). Thus, we will not be concerned with connectionist research on the enormously complex issues involved in dealing with the complexity, variability, and noisiness of acoustic representations of speech (see, e.g., Salmela, Lehtokangas, & Saarinen, 1999, for a typical application of connectionist methods to speech technology). We also shall not deal with higher-level aspects of linguistic processing. Nonetheless, as we shall see, even given these restrictions, the problem of understanding human speech processing is still formidable.

Naïvely, we might imagine that the speech processor has to do two jobs, one after the other. First, it has to segment speech input into units corresponding to words (i.e., it has to find word boundaries); second, it has to recognize each word. But on reflection, this viewpoint seems potentially problematic, because it is not clear how the speech processor can determine where the word boundaries are until the words are recognized. And conversely, word recognition itself seems to presuppose knowing which chunk of speech material corresponds to a potential word. Thus, segmentation and recognition appear to stand in a chicken-and-egg relationship—each process seems to depend on the other.

One approach to resolving the paradox is to assume that segmentation and recognition are two aspects of a single process, that tentative hypotheses about each issue are developed and tested simultaneously, and mutually consistent hypotheses are reinforced. A second approach is to suppose that there are segmentation cues in the input that are used to give at least better-than-chance indications of what segments may correspond to identifiable words. So the question is this: Does speech processing involve dedicated segmentation strategies prior to word recognition?

Developmental considerations suggest that there may be specialized segmentation methods. The infant, initially knowing no words, seems constrained to segment speech input using some method not requiring word recognition. Moreover, infant studies have shown that prelinguistic infants may use such methods, and are sensitive to a variety of information that is available in the speech stream and potentially useful for segmentation, such as phonotactics and lexical stress (Jusczyk, 1997).

Connectionist models have begun to address questions of how effective different kinds of segmentation cues might be. For example, Cairns, Shillcock, Chater, and Levy (1997) explore a model of segmentation based on predictability. They note that language is less predictable across, rather than between, words. They trained a recurrent network on a large corpus of phonologically transcribed conversational speech, represented as a sequence of bundles of binary phonetic features. The network was trained to predict the next bundle of features along with the previous and current feature bundles, based on the current input material. Where prediction error was large, it was assumed that a word boundary had been encountered. This model captured some aspects of human segmentation performance. For example, it spontaneously learned to pay attention to patterns of strong and weak syllables as a segmentation cue. However it was able to reliably predict only a relatively small proportion of word boundaries, indicating that other cues also need to be exploited. While the Cairns et al. model uses just a single cue to segmentation, Christiansen, Allen, and Seidenberg (1998) showed how multiple, partial constraints on segmentation could yield much better segmentation performance. They trained an SRN to integrate sets of phonetic features with information about lexical stress (strong or weak) and utterance boundary information (encoded as a binary unit) derived from a corpus of child-directed speech. The network was trained to predict the appropriate values of these three cues for the next segment. After training, the network was able to integrate the input such that it would activate the boundary unit not only at utterance boundaries, but also at word boundaries inside utterances. The network was thus able to generalize patterns of cue information that occurred at the end of utterances to cases where the same patterns occurred within an utterance. This model performed well on the word-segmentation task while capturing additional aspects of infant segmentation, such as the bias toward the dominant trochaic (strong–weak) stress pattern in English, the ability to distinguish between phonotactically legal and illegal novel words, and having segmentation errors being constrained by English phonotactics.

This model shows how integrating multiple segmentation cues can lead to good segmentation performance. To what extent does it provide a model of how infants process speech? Christiansen, Conway, and Curtin (2000) used the trained model, without any additional modifications, to fit recent infant data. These data are of particular interest, because they have been claimed to be incompatible with a purely connectionist approach to language processing, and to require the language processor to use "algebraic" or symbolic rules (Marcus, Vijayan, Rao, & Vishton, 1999). Marcus et al. habituated infants on syllable sequences that followed either an AAB or ABB pattern (e.g., *le-le-je* versus *le-je-je*). The infants were then presented with sequences of novel syllables, either consistent or inconsistent with the habituation pattern, and showed a preference for the inconsistent items. Christiansen et al. suggested that statistical knowledge acquired in the context of learning to segment fluent speech provided the basis for these results, in much the same way as knowledge acquired in the process of learning to read can be used to perform experimental tasks such as lexical decision. Their simulation closely replicated the experimental conditions, using the same number of habituation and test trials as in the original experiment (no repeated training epochs) and one network for each infant. Analyses of the model's segmentation performance revealed that the model was significantly better at segmenting out the syllables in the inconsistent items. This makes the inconsistent items more salient and therefore explains why the infants preferred these to the consistent items. Thus, Christiansen et al.'s results challenge the claim that the Marcus et al. infant data necessarily require that the infant's language-processing system is using algebraic rules. Moreover, these infant data provide an unexpected source of evidence for the Christiansen et al. model, viewed as a model of infant segmentation.

Segmentation cues are potentially important in guiding the process of word recognition. But even if such cues are exploited very effectively, segmentation cues alone can achieve only limited results. A definitive segmentation of speech can only occur after word recognition has occurred. Speech is frequently locally ambiguous: To use an oft quoted example, it is difficult to distinguish "recognize speech" from "wreck a nice beach" when these phrases are spoken fluently. These interpretations correspond to very different segmentations of the input. It is therefore clear that bottom-up segmentation cues alone will not always segment the speech stream into words reliably. In such cases of local ambiguity, a decisive segmentation of the input can only be achieved when the speaker has recognized which words have been said. This theoretical observation ties in with empirical evidence that strongly indicates that during word recognition in adulthood multiple candidate words are activated, even if these correspond to different segmentation of the input. For example, Gow and Gordon (1995) found that adult listeners hearing sentences involving a sequence (e.g., *two lips*) that could also be a single word (*tulips*) showed speeded processing of an associate of the second word (*kiss*) and to an associate of the longer word (*flower*), indicating that the two conflicting segmentations were simultaneously entertained. This would not occur if a complete segmentation of the input occurred before word recognition was attempted. On the other hand, it is not clear how these data generalize to word segmentation and recognition in infancy before any comprehensive vocabulary has been established. How segmentation and recognition develop into the kind of integrated system evidenced by the Gow and Gordon data remains a matter for future research.

Gow and Gordon's (1995) result also suggests that word recognition itself may be a matter of competition between multiple activated word representations, where the activation of the word depends on the degree of match between the word and the speech input. Indeed, many studies point toward this conclusion, from a range of experimental paradigms. Such competition is typically implemented in connectionist networks by a localist code for words (the activation of a single unit represents the strength of evidence for that word, with inhibitory connections between word units). Thus, when an isolated word is identified, a "cohort" of words consistent with that input is activated; as more of the word is heard, this cohort is rapidly reduced, perhaps to a single item.

While competition at the word level has been widely assumed, considerable theoretical dispute has occurred over the nature of the interaction between different levels of mental representation. Bottom-up (or "data-driven") models are those in which less abstract levels of linguistic representation feed into, but are not modified by, more abstract levels (e.g., the phoneme level feeds to the word level, but not the reverse). We note, however, that this does not prevent these models from taking advantage of suprasegmental information, such as in the inclusion of lexical stress in the Christiansen et al. (2000) segmentation model, provided that this information is available in a purely bottom-up fashion (i.e., no lexical-level feedback). Interactive (also "conceptually-driven" or top-down) models allow a two-way flow of information between levels of representation. Figures 2.1 and 2.4 provide abstract illustrations of the differences in information flow between the two types of models. Note that bottom-up models allow information to flow through the network in one direction only, whereas interactive models allow information to flow in both directions.

The bottom-up versus interactive debate rages in all areas of language processing, and also in perception and motor control (e.g., Bruner, 1957; Fodor, 1983; Marr, 1982; Neisser, 1967). Here we focus on putative interactions between information at the phonemic and the lexical levels in word recognition (i.e., between phonemes and words), where experimental work and connectionist modeling has been intense.

The most obvious rationale for presuming that there is top-down information flow from the lexical to the phoneme level stems from the effects of lexical context on phoneme identification. For example, Ganong (1980) showed that the identification of a syllable-initial speech sound, constructed to be between a /g/ and a /k/, was influenced by lexical knowledge. This intermediate sound was predominantly heard as a /k/ if the rest of the word was -iss (*kiss* was favored over *giss*), but heard as /g/ if the rest of the word was -ift (*gift* was favored over *kift*).

The early and very influential TRACE model of speech perception (McClelland & Elman, 1986) attempts to explain data of this kind from an interactive viewpoint. The model employs the standard interactive activation network architecture already described, with layers of units standing for phonetic features, phonemes, and words. There are several copies of each layer of units, standing for different points in time in the utterance, and the number of copies differs for each layer. At the featural level, there is a copy for each discrete "time slice" into which the speech input is divided. At the phoneme level, there is a copy of the detector for each phoneme centered over every three time slices. The phoneme detector centered on a given time slice is connected to feature detectors for that time slice, and also to the feature detectors for the previous three and subsequent three slices. Hence, successive detectors for the same phoneme overlap in the feature units with which they interact. Finally, at the word level there is a copy of each word unit at every three time slices. The window of phonemes with which the word interacts corresponds to the entire length of the word. Here, again, adjacent detectors for the same word will overlap in the lower-level units to which they are connected. In short, then, we have a standard interactive activation architecture, with an additional temporal dimension added, to account for the temporal character of speech input. TRACE captures the Ganong effect because phoneme and lexical identification occur in parallel and are mutually constraining. TRACE also captures experimental findings concerning various factors affecting the strength of the lexical influence (e.g., Fox, 1984), and the categorical aspects of phoneme perception (Massaro, 1981; Pisoni & Tash, 1974). TRACE also provides rich predictions concerning the time course of spoken word recognition (e.g., Cole & Jakimik, 1978; Marslen-Wilson, 1973; Marslen-Wilson & Tyler, 1975), and lexical influences on the segmentation of speech into words (e.g., Cole & Jakimik, 1980).

TRACE provides an impressive demonstration that context effects can indeed be modeled from an interactive viewpoint. But context effects on phoneme recognition can also be explained in purely bottom-up terms. If a person's decisions about phoneme identity depend on both the phonemic and lexical levels, then phoneme identification may be lexically influenced, even though there need be no feedback from the lexical to the phoneme level. For example, the Ganong effect might be explained by assuming that