CHAPTER 17

Computational Models of Psycholinguistics

Nick Chater and Morten H. Christiansen

1. Introduction

The computational mechanisms that underlie how people process and acquire language has been a central topic for cognitive science research since the beginning of the field. Indeed, Chomsky's revolutionary impact on linguistics (e.g., 1957, 1959, 1965) involved the attempt to align linguistics with the project of cognitive science. The project of linguistics was viewed as providing a formally specified account of the knowledge that underpins linguistic behavior. This specification took the form of a generative grammar – a set of rules that determined which linguistic forms (strings of phonemes, strings of words, etc.) are linguistically acceptable and which are not. Generative grammars themselves had direct relationships to models of formal languages in automata theory and are used to specify formal languages, both in logic and the development of programming languages.

For Chomsky, computational ideas were also fundamental to understanding human language in another way. He defined a formal hierarchy of grammars and associated languages (regular, context-free, context-sensitive, unrestricted grammars), each of which relates elegantly to the kind of language-processing operations that can parse and produce them (Chomsky & Schützenberger, 1963). Thus, for example, a finite state automaton can parse and produce only finite-state languages, a push-down automaton can deal with finite-state and context-free languages, and so on. Moreover, Chomsky used these observations to devastating effect, in considering existing behaviorist theories of linguistic behavior (Chomsky, 1959; Skinner, 1957). He argued that human languages correspond to the highest level in the Chomsky hierarchy and, hence, cannot be accounted for by existing associative theories, which appear to be limited to processing mechanisms that correspond to a finite state machine. Indeed, Chomsky's arguments concerning the formal and computational properties of human language were one of the strongest and most influential lines of argument behind the development of the field of cognitive science, in opposition to behaviorism. Moreover, Chomsky's (1968, 1980) arguments

concerning the poverty of the linguistic input available to the child in relation to the spectacular intricacy of the linguistic system that children acquire became a major impetus for strongly nativist theories of language (e.g., Berwick, 1986; Crain, 1991; Lightfoot, 1991; Pinker, 1984), but also, by extension, nativist theories across a wide range of cognitive domains (e.g., Hirschfeld & Gelman, 1994; Pinker, 1997).

Given this historical background, it is perhaps not surprising that computational models of language processing and acquisition have been theoretically central to the development of cognitive science over the past fifty years. But the direction that these models have taken has been much less predictable. Chomsky's initial suggestion that a formal theory of linguistic knowledge should integrate smoothly with computational theories of processing and acquisition has run into a number of difficulties. Reactions to these difficulties vary widely, with the result that computational models in psycholinguistics have fragmented into different traditions, which are not readily integrated into a single perspective. The resulting work has been rich and varied, and has led to considerable qualitative insights into many aspects of human language processing and acquisition; but it is by no means clear how to synthesize the variety of computational methods and insights into anything resembling an integrated theoretical framework. This chapter outlines the historical origins and the state of the art of computational models of psycholinguistic processes. Also considered are the interrelationships between the different theoretical traditions that have emerged from, and in reaction to, the Chomskyan revolution. This survey is necessarily highly selective, both in terms of the topics covered and the research within each topic. The survey aims, though, to focus attention on topics that have the widest general theoretical implications, both for other fields of computational cognitive modeling and for the project of cognitive science more broadly.

The next section, *Three Computational Frameworks for Psycholinguistics*, begins by outlining and contrasting symbolic, connectionist, and probabilistic approaches to the computational modeling of psycholinguistic phenomena (see Chapter 1 in this volume). There are important overlaps and relationships between these traditions, and each tradition itself contains a range of incompatible viewpoints. Nonetheless, this three-way division is at least a convenient starting point for discussion. Next, attention turns to specific computational proposals and associated theoretical positions across specific psycholinguistic topics. *From Signal to Word* considers word segmentation and recognition, and single word reading. *Sentence Processing* primarily focusses on parsing, relating connectionist and probabilistic models to the symbolic models of grammar and processing associated with Chomsky's program. *Language Acquisition* reviews formal and computational models of language learning and re-evaluates, in the light of current computational work, Chomsky's early theoretical arguments for a strong nativist view of the computational mechanisms involved. Finally, in *Where Next?* the future of computational models of psycholinguistics is considered.

2. Three Computational Frameworks for Psycholinguistics

2.1. *Chomsky and the Symbolic Tradition*

Chomsky's initiation of the cognitive science of language proposed that human language should be assimilated into the domain of formal languages, and this immediately suggests that the computational mechanisms involved in parsing and producing formal languages, which is a rich area of research in computer science, (e.g., Aho & Ullman, 1972; Hopcroft, Motwani, & Ullman, 2000), might be co-opted and extended to provide models of human language processing. This is a rigorously *symbolic* perspective on the structure of language and the nature of the computational processes operating over language – a perspective that meshed well with the prevalent computational models of mind, inspired by spectacular

theoretical and technical advances in symbolic computation (e.g., Winograd, 1972).

This perspective provides an attractively crisp picture of the relationship between knowledge of the language, and the processing operations used in parsing or producing it. The knowledge of the language is embodied in a set of declarative rules (i.e., the rules are explicitly represented in symbolic form), and a set of processing operations applies these rules in parsing and production. In parsing, the problem is to find a syntactic derivation (typically corresponding to a tree structure), using the rules, that yields the observed sequence of words; in production, there is the converse problem of using the rules to construct a derivation and then to output the resulting sequence of words. From this point of view, too, the problem of language learning can be stated as a problem of induction, that is, inducing a grammar (i.e., a set of symbolic linguistic rules) from a set of observed sentences, and this problem yields readily to formal analysis, using techniques from theoretical computer science.

Yet, despite these evident strengths, and moreover, extensive developments in linguistic theory based on the symbolic approach, the expected program of computational models of psycholinguistic phenomena rapidly ran into difficulties. Initial grammar formalisms proposed that the derivation of a sentence required the operation of a succession of transformations, leading to the natural assumption that a computational model of language parsing and production would need to recapitulate these transformations and that the number and complexity of transformations should therefore correlate with processing time and difficulty in psycholinguistic experiments. This derivational theory of complexity (Miller, 1962) proved to be a poor computational model when compared with empirical data and was rapidly abandoned. In the generative grammar tradition, the relationship between linguistic theory and processing was assumed to be indirect (e.g., Fodor, Bever, & Garrett, 1974), and this led subsequent developments in the Chomskyan tradition in generative grammar to disengage from work on computational modeling.

Yet, in parallel with this, a wide range of research in computational linguistics took the generative approach to linguistic theory and attempted to build computational mechanisms for language processing that could serve as potential cognitive models. For example, early debates concerned alternative mechanisms for parsing versions of transformational grammar or related but simpler formalisms, for example, Wanner and Maratsos's (1978) Augmented Transition Networks and Frazier and Fodor's (1978) "sausage machine." Work on cognitive models of symbolic parsing has also continued (e.g., Crocker, 1996; Gibson, 1998). Early and recurring issues arising from these models concerned how to deal with the huge local ambiguity of natural language, which appears to lead to a combinatorically explosive number of possible parses. Are many parallel parses computed at once? If not, what constraints determine which parse is pursued? (Marcus, 1980).

Psycholinguistic theories focusing on the generative tradition tended to assume that language processing is an autonomous domain (Ferreira & Clifton, 1986), that is, language processing can be separated from processes of general world knowledge (Fodor, 1983). Moreover, it is typically assumed that structural, rather than probabilistic, features of language are central. The idea is that the cognitively represented linguistic rules determine what it is possible to say (the linguistic rules aim to capture linguistic *competence*; Chomsky, 1965); all manner of pragmatic, and knowledge-based constraints, as well as processing limitations, will determine what people actually say (such matters are assumed to be theoretically secondary issues of *performance*, Chomsky, 1965). For these reasons, early proposals concerning parsing and production assumed these processes to be determined by aspects of syntactic structure, that is, that the processing system may aim to build a tree with as few nodes as possible (the core of Frazier's [1979] proposal of minimal attachment).

Attempts to model psycholinguistic data have, however, been relatively rare, in the symbolic tradition. Purely structural features of language appear to be just one of the factors that determine performance in psycholinguistic experiments, for example. Predictably enough, experimental results are strongly influenced by the very probabilistic and world-knowledge factors that the Chomskyan viewpoint aims to relegate to the realm of "performance" (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell & Tanenhaus, 1992). One approach is to view psycholinguistic paradigms as highly imperfect measures of the "pure" language module – and, indeed, classical linguistics has typically taken this perspective and ignored experimental psycholinguistic evidence concerning the structure of language to the exclusion of direct linguistic acceptability judgments from native speakers. Another approach is to propose that structural factors determine which options the processor considers and that probability and world knowledge may arise purely in pruning such proposals (Crain & Steedman, 1985; Fodor, 1983). From the point of view of building computational models of psycholinguistic processes, it seems inevitable that probabilistic aspects of language processing must take center stage. This observation has been one line of impetus behind two rather different, but related, alternative computational frameworks: connectionist and probabilistic models of language, which are discussed in the following subsections.

2.2. *Connectionist Psycholinguistics*

A symbolic perspective on language processing fits well with, and was a strong motivation for, the broader view of the mind as a symbol-processing system, based on principles analogous to the digital computer (Newell & Simon, 1972). Connectionism has a different origin in attempts to design computers inspired by the brain (see Chapter 2 in this volume). At a coarse level, the brain consists of a very large number of densely interconnected neurons, each of which is computationally relatively simple. These neurons do not appear to operate individually in tackling information processing problems; rather, large numbers of neurons operate simultaneously and co-operatively to process information. Furthermore, neurons appear to communicate real numbers (approximately encoded by firing rate) rather than symbolic messages, and therefore neurons can be viewed as mapping real-valued inputs (from other neurons) onto a real-valued output (which is transmitted to other neurons). Connectionist nets mimic these properties, although typically without attempting high levels of biological realism (although see Dayan & Abbott, 2001). Connectionist methods also provide interesting "bottom-up" models of learning – learning occurs by a multitude of small adjustments to the "weights" of the connections between processing units, which can be determined purely locally; this is a very different picture of learning from the traditional serial hypothesis-generation and test envisaged by typical symbolic models of learning. This raises the possibility that connectionism may shed new light on processes underlying language acquisition (Bates & Elman, 1993; Elman 1993, 2003; Redington & Charter, 1998).

The relative merits of connectionist and symbolic models of language are, as noted earlier, hotly debated. But should they be in competition at all? Advocates of symbolic models of language processing assume that symbolic processes are somehow implemented in the brain: They, too, are connectionists, at the level of *implementation*. They assume that language processing can be described both at the psychological level, in terms of symbol processing, and at an implementational level, in neuroscientific terms (to which connectionism approximates). If this is right, then connectionist modeling should start with symbol processing models of language processing and implement these in connectionist nets. Advocates of this view (Fodor & Pylyshyn, 1988; Marcus, 1998; Pinker & Prince, 1988) typically assume that it implies that symbolic modeling is entirely autonomous from connectionism;

symbolic theories set the goalposts for connectionism, but not the reverse. Chater and Oaksford (1990) argued that, even according to this view, there will be a two-way influence between symbolic and connectionist theories, since many symbolic accounts can be ruled out precisely because they could not be neurally implemented to run in real time. Indeed, some computational proposals concerning, for example, morphology or reading single words, have a hybrid character, in which aspects of what is fundamentally a symbolic process are implemented in connectionist terms, to explain, for example, complex statistical patterns in dealing with irregular items, alongside apparently rigid rule-based patterns, for regular items (e.g., Coltheart et al., 2001; Marcus, 2000).

Many connectionists in the field of language processing have a more radical agenda: to challenge, rather than reimplement, the symbolic approach. They see many aspects of language as consisting of a multitude of "soft" regularities, more naturally captured by connectionist, rather than rule-based, methods (e.g., Seidenberg, 1997). There are also theoretical positions that take inspiration from both symbolic and connectionist paradigms: In linguistics, optimality theory attempts to define a middle ground of ranked, violable linguistic constraints, used particularly to explain phonological regularities (Smolensky & Legendre, 2006). And in morphology, there is debate over whether "rule + exception" regularities (e.g., English past tense, German plural) are better explained by a single stochastic process (Hahn & Nakisa, 2000; Marcus et al., 1995). Overall, then, a central theoretical question is how far connectionist models complement, or compete with, symbolic models of language processing and acquisition (Marcus, 1998; Seidenberg & Elman, 1999; Smolensky, 1999; Steedman, 1999).

2.3. *Probabilistic Models of Language*

As noted earlier, according to Chomsky (1965), the study of language should primarily focus on *competence*, rather than performance, that is, what is linguistically acceptable, rather than the statistical properties of what people actually say. This has led to the downplaying of probabilistic features of language, more generally, in favor of the putatively rigid linguistic rules (although there has been a long tradition of interest in statistical properties of language in sociolinguistics, e.g., Labov, 1972).

Yet, recent work, particularly in computational linguistics and, as is described later, connectionist psycholinguistics, has suggested that a probabilistic viewpoint may be central to understanding language processing, language acquisition, and perhaps the structure of language itself (Chater & Manning, 2006). Thus, for example, whereas from a symbolic perspective, parsing is naturally viewed as the problem of constructing a logical derivation from grammatical rules to a string of words generated by the application of those rules (Pereira & Warren, 1983), from a probabilistic point of view, the problem is not merely to find *any* derivation, but to find the *most probable* derivation (or the most probable derivations, ranked by their probability). Moreover, given the notorious local ambiguity of language (where large numbers of lexical items are syntactically ambiguous and can combine locally in many ways), focusing on the most probable local derivation can potentially lead to a dramatic reduction in the problem of searching for globally viable parses.

In particular, probabilistic Bayesian methods (see Chapter 3 in this volume) specify a framework showing how information about the probability of generating different grammatical structures and their associated word strings can be used to infer grammatical structure from a string of words. An elegant feature of the probabilistic viewpoint is that the same Bayesian machinery can also be turned to the problem of learning: of showing how information about the degree to which different probabilistic grammars have different probabilities of generating observed linguistic data and using this to infer grammars, at least to a limited extent, from linguistic data. Moreover, this Bayesian framework is analogous to probabilistic models of vision, inference,

and learning (Chater, Tenenbaum, & Yuille, 2006); what is distinctive is the specific structures (e.g., syntactic trees, dependency diagrams) relevant for language.

As with the relationship between symbolic and connectionist viewpoints, the relationship between probabilistic and symbolic views can be viewed as complementary or competitive. The complementary viewpoint needs assume only that probabilities are *added* to existing linguistic rules to indicate how often rules of each type are used; a clean separation between nonprobabilistic linguistic competence and probabilistic information and processing used in linguistic *performance* can thus be maintained. But the more radical viewpoint is that some, and perhaps many, aspects of language structure should be viewed probabilistically (Bod, Hay, & Jannedy, 2003).

In linguistics, there has been renewed interest in phenomena that seem inherently graded and/or stochastic, from phonology to syntax (Fanselow et al., in press; Hay & Baayen, 2005). There have also been revisionist perspectives on the strict symbolic rules thought to underlie language and an increasing emphasis on nonrule-based processes, for example, processes based on individual linguistic constructions (Goldberg, 2006; Tomasello, 2003). Indeed, some theorists suggest that many aspects of language processing and acquisition may be best understood in terms of retrieving similar previous cases from a large store of prior instances of linguistic structure (Bod, 1998; Daelemans & van den Bosch, 2005). Memory, or instance-based, views are currently widely used across many fields of cognitive science.

3. From Signal to Word

Early theories of speech processing adopted a symbolic viewpoint in which a set of symbolically represented word forms were matched against the acoustic or visual input, in some cases, assuming a sequential search in memory, by analogy with the operation of memory retrieval in digital computers (Forster, 1976). Other early models assumed that multiple word forms could be activated in parallel, and choice was resolved by a process of competition (Morton, 1969); and, in the context of speech, this competition was assumed to proceed incrementally, and very rapidly, as the speech signal was encountered (Marslen-Wilson & Welsh, 1978).

These models were typically not implemented, however, and hence not quantitatively matched against empirical data. Two sources of candidate computational models began to emerge, however. The first arose from the application of sophisticated probabilistic and mathematical techniques, such as from hidden Markov models, vector quantization, and dynamic programming, in the development of speech technology (Juang & Rabiner, 1991). These technical developments had relatively little impact on psychological theories of speech processing, although the probabilistic tradition that they embody has more recently had a substantial impact, as will become clear. The second source of candidate models arose from connectionism, which led to a range of important detailed cognitive models. Connectionist modeling of speech processing begins with TRACE, which has an "interactive activation" architecture, with a sequence of "layers" of units (Figure 17.1A), for phonetic features, phonemes, and words (McClelland & Elman, 1986). Speech input corresponds to activation of phonetic features, which allow the recognition of phonemes and the words; at each level, representations compete via mutually inhibitory links. Hence, alternative phonemes compete to explain particular bundles of phonetic features, and different hypothetical words inhibit each other. Between layers, mutually consistent hypotheses support each other, for example, phonetic features consistent with a particular phoneme reinforce each other; a word and its constituent phonemes are mutually reinforcing. The bidirectional, interactive character of these links underpins TRACE's ability to capture apparently top-down effects – if there is evidence that a particular word has been heard, that word will support a constituent phoneme for which the input at the phonetic level might be ambiguous.

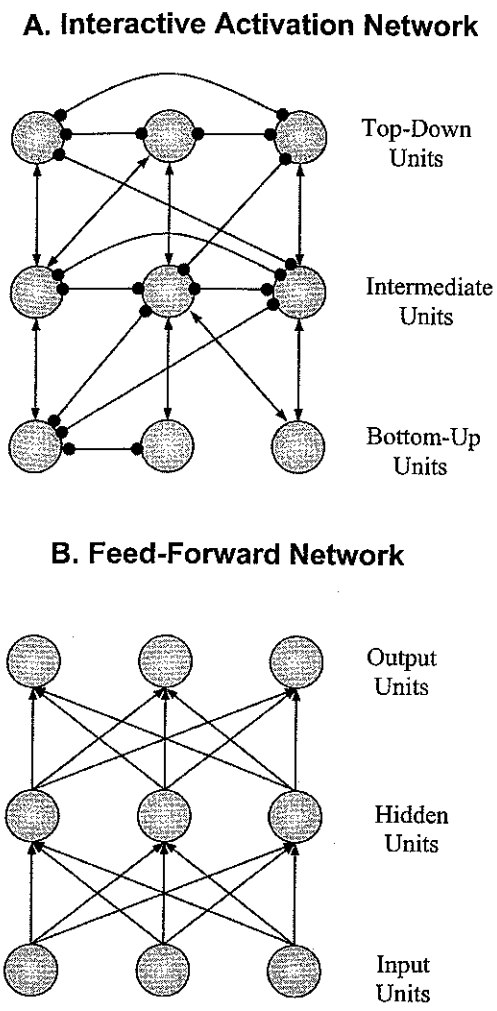


Figure 17.1. Interactive and feed-forward connectionist architectures. A fundamental divide between neural network architectures concerns whether the input is processed unidirectionally or whether top-down feedback is allowed. (A) Top-down feedback is a distinctive feature of the *interactive activation network* (as used in TRACE; McClelland & Elman, 1986). The network has bidirectional excitatory (arrows) or inhibitory (filled circles) links. Activation flows bottom-up and top-down, reinforcing mutually consistent states and inhibiting inconsistent states. Inhibitory connections within layers implement competition between alternative words, phonemes, or phonetic features. The weights in TRACE are hand-coded rather than learned. (B) A *feedforward network* passes information in one direction, with no feedback connections. Feedforward networks are typically not hand-coded, but are trained using backpropagation, which minimizes the discrepancy between the network's actual and desired output. Information flows bottom-up from input to output units (see Chapter 2 in this volume).

TRACE captured a wide range of empirical data and made important novel predictions. TRACE is most controversial because it is interactive – the bidirectional links between units mean that information flows top-down as well as bottom-up. Other connectionist models, by contrast, assume purely bottom-up information flow (Norris, 1994). TRACE provided an impetus for the interactive versus bottom-up debate, with a prediction apparently incompatible with bottom-up models.

To understand this novel prediction, it is necessary to sketch two background results on speech perception. First, note that if an ambiguous phoneme is sometimes resolved by the word context in which that phoneme is embedded. Thus, if an ambiguous /s/-/š/ phoneme (the /s/ and /š/ are pronounced as in the onsets of the words *sip* and *ship*) is presented at the end of fooli-, it is heard as a /š/, because *foolish* is a word and *fooliss* is not; conversely, in the context *Christma-*, the same ambiguous phoneme is heard as a /s/, because *Christmas* is a word, whereas *Christmash* is not. This is the Ganong effect (Ganong, 1980), and it follows naturally from an interactive viewpoint, where word recognition can feed back to phoneme recognition. But it can be equally well explained by bottom-up models by assuming that the participant's responses concerning phoneme identity are simultaneously influenced by both the phoneme and lexical levels (Fodor, 1983). The second background result to motivate Elman and McClelland's (1988) prediction comes from the observation that, in natural speech, the pronunciation of a phoneme is affected by surrounding phonemes: this is "coarticulation." Thus, for example, /t/ and /k/ differ only by the phonetic feature of place of articulation, that is, tongue position. But the location of the tongue for the current phoneme is also influenced by its previous position and hence by the previous phoneme. In particular, for example, /s/ and /t/ have the same place of articulation; but after a /š/, the place of articulation of the /t/ is dragged somewhat toward that which is normal for a /k/. The opposite pattern occurs for /k/, which is dragged somewhat toward the pronunciation of a /t/

when preceded by a /s/. Mann and Repp (1981) put an ambiguous /k/-/t/ phoneme in the context of *–apes*, so that when heard alone, the input was judged equally often to the *capes* or *tapes*. After the word *Christmas*, the ambiguous input is most often heard as *capes* – the ambiguous phoneme is "explained" by the speech processor by the influence of the previous /š/ context. Conversely, after the word *foolish*, the same ambiguous phoneme is heard as *tapes*. Thus, Mann and Repp (1981) concluded that the speech processor engages in "compensation for coarticulation" (CFC); it compensates for the coarticulation that arises in speech production.

Elman and McClelland (1988) observed that TRACE makes an interesting prediction where the preceding phoneme is also ambiguous – between /š/ and /s/. If the word level directly influences the phoneme level (and this type of direct interactive influence is what leads to the Ganong effect), then the compensation of the /k/ should occur even when the /s/ relies on lexical input for its identity (i.e., with an ambiguous /s/-/š/ in *Christmas*, the /s/ should be restored and thus CFC should occur as normal, so that the following ambiguous /k/-/t/ should be perceived as /k/). TRACE's novel prediction was experimentally confirmed (Elman & McClelland 1988).

3.1. *Bottom-Up Connectionist Models Capture "Top-Down" Effects*

Yet, bottom-up connectionist models *can* capture these results. One study used a simple recurrent network (SRN; Elman, 1990) to map phonetic input onto phoneme output (Norris, 1993). The SRN is a standard feedforward network (Figure 17.1B), where hidden units are copied back at a given time-step and presented to the network at the next time-step, so that the network's behavior is determined by a sequence of inputs, rather than just the current input (Figure 17.2A). In Norris's model, when the SRN received phonetic input with an ambiguous first word-final phoneme and ambiguous initial segments of the second word, an analog of CFC was observed. The percentages of /t/ and /k/ responses to the first phoneme of the second word depended on the identity of the first word (as in Elman & McClelland, 1988). Importantly, the explanation for this pattern of results cannot be top-down influence from word units, because there *are* no word units. Nonetheless, the presence of "feedback" connections in the hidden layer of the SRN might suggest that some form of interactive processing occurs in this model. But this is misleading – the feedback occurs within the hidden layer (i.e., from its previous to its present state), rather than flowing from top to bottom. This model, although an important demonstration, is small-scale – it deals with just twelve words. However, a subsequent study scaled up these results using a similar network trained on phonologically transcribed conversational English (Cairns et al., 1997).

How is it possible that bottom-up processes can mimic what appear to be top-down effects from the lexicon? It was argued that restoration depends on local statistical regularities between the phonemes within a word, rather than depending on access to lexical representations, thus, individual phonemes are supported by phonemes in the same word, not via links to an abstract word-level representation but instead by lateral connections between the phonemes, exploiting the statistical dependencies between neighboring phonemes. More recent experiments have since shown that CFC is indeed determined by statistical regularities for nonword stimuli, and that, for word stimuli, there appear to be no residual effects of lexical status once statistical regularities are taken into account (Pitt & McQueen, 1998). It is not clear, though, whether bottom-up models can model other evidence that phoneme identification is affected by the lexicon, for example, from signal detection analyses of phoneme restoration (Kraljic & Samuel, 2005; Norris, McQueen, & Cutler, 2000; Samuel, 1996).

3.2. *Exploiting Distributed Representations*

A different line of results provides additional evidence that bottom-up models

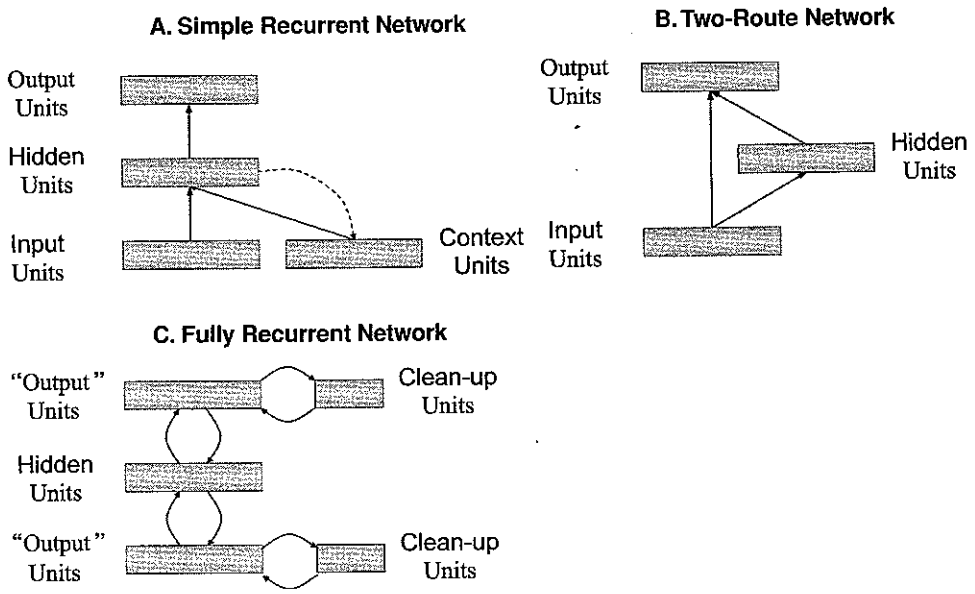


Figure 17.2. Connectionist architectures for cognitive models. The structure of the network is represented more schematically than in Figure 17.1. Each block represents a bank of units, and arrows between blocks indicate full connectivity between each unit in the relevant blocks. (A) A *simple recurrent network* is essentially a standard feedforward network equipped with an extra layer of so-called context units. At each-time step, an input propagates through the hidden units to the outputs (solid arrows). The hidden-unit activation at the previous time-step is copied back to the context layer (dashed arrows) and paired with the current input (solid arrows). Thus, the hidden units influence the processing of subsequent inputs, providing a limited ability to deal sequential inputs. (B) Here, there are two routes from input to output – one that is direct and can only encode simple relationships and one that is indirect and can encode more complex relationships between input and output. Zorzi, Houghton, & Butterworth (1998) use this structure to illustrate how a single network can simultaneously learn both simple patterns (the basic grapheme-phoneme correspondences of English) using one route while simultaneously learning a complex pattern of exceptions with the other route. Note that the resulting model of reading differs from standard dual-route models (Coltheart et al., 2001) because the route encoding lexical exceptions can operate by "correcting" the simpler route rather than proceeding independently of it. (C) A fully recurrent network in which activation flows both up and down the network and also recirculates between banks of units (Allen & Seidenberg, 1999). The network can be trained using the backpropagation-through-time learning algorithm (e.g., Williams & Zipser, 1990). The recirculation within banks of units serves to "clean up" any errors that may have been introduced. The labels "input" and "output" are shown in quotes because all connections are bidirectional. Thus, although the network can be used to map form to meaning (as in Allen & Seidenberg, 1999), it could equally be used to map meaning to form.

can accommodate apparently top-down effects (Gaskell & Marslen-Wilson, 1995). An SRN was trained to map a systematically altered featural representation of speech onto a phonemic and semantic representation of the same speech (following previous work, Kawamoto, 1993). After training, the network showed evidence of lexical effects in modeling lexical and phonetic decision data (Marslen-Wilson & Warren, 1994). This work was extended by an SRN trained to map sequential phonetic input onto corresponding distributed representations of phonological surface forms and semantics (Gaskell & Marslen-Wilson, 1997a, 1997b). This style of representation

contrasts with the localist representations used in TRACE. The ability of the SRN to model the integration of partial cues to phonetic identity and the time course of lexical access provides support for a distributed approach. An important challenge for such distributed models is to model the simultaneous activation of multiple lexical candidates necessitated by the temporal ambiguity of the speech input (e.g., /kæp/ could continue*captain* and *captive*; see Allopenna, Magnuson, & Tanenhaus, 1998, for a generalization of this phenomenon). The "coactivation" of several lexical candidates in a distributed model results in a semantic "blend" vector. Computational explorations (Gaskell & Marslen-Wilson, 1999) of such semantic blends provide explanations of recent empirical results aimed at measuring lexical coactivation (Gaskell & Marslen-Wilson, 1997a, 1997b) and more generally provide a concrete implementation of theoretical proposals that were previously expressed informally (Marslen-Wilson & Warren, 1994).

3.3. *Speech Segmentation*

Further evidence for the bottom-up approach to speech processing comes from the modeling of speech segmentation (Christiansen, Allen, & Seidenberg, 1998). An SRN was trained 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 generalize patterns of cue information that occurred at the end of utterances to when the same patterns occurred elsewhere in the utterance. Relying entirely on bottom-up information, the model performed well on the word segmentation task and captured important aspects of infant speech segmentation. Speech segmentation has also been the subject a wide variety of alternative computational proposals (e.g., Hockema, in press).

3.4. *Reading Aloud*

Connectionist research on reading aloud has focused on single words. A classic early model used a feedforward network (Figure 17.1B) to map from a distributed orthographic representation to a distributed phonological representation for monosyllabic English words (Seidenberg & McClelland, 1989). The net's performance captured a wide range of experimental data on the assumption that network error maps onto response time. This model contrasts with standard views of reading, which assume both a "phonological route," applying rules of pronunciation, and a "lexical route," which is a list of words and their pronunciations. Regular words (e.g., *sing*) can be read using either route, exception words (e.g., *colonel*) by the lexical route, and nonwords by the phonological route. It was claimed that, instead, a single connectionist route can pronounce both exception words and nonwords. Critics have responded that the network's nonword reading is well below human performance (Besner et al., 1990, although see Seidenberg & McClelland, 1990). Another difficulty is the model's reliance on (log) frequency compression during training (otherwise exception words are not learned successfully). Subsequent research has addressed both limitations, showing that a network trained on actual word frequencies can achieve human levels of performance on both word and nonword pronunciation, which has led to a range of new connectionist models (e.g., Plaut, 1999).

3.5. *Explaining the Acquired Dyslexias*

The number of routes by which words can be recognized is a central point of theoretical debate. It is widely agreed that both semantic (where orthography is mapped to meaning and then to phonology) and nonsemantic routes (which map orthography to phonology without going through semantics) are available. The key controversy is whether there are one or two *non-semantic* routes. Dual-route theorists

typically argue that there are two such routes – a phonological route that uses the rules of regular orthography/phonology to pronounce words piecemeal and a "lexical" route, which maps whole orthographic inputs to whole phonological outputs by a process akin to table look-up. Some connectionists take a single nonsemantic route viewpoint, arguing, for example, that the division of labor between phonological and semantic routes can explain diverse neuropsychological syndromes that have been taken to require a dual-route account (Plaut et al., 1996). One viewpoint is that a division of labor emerges between the phonological and the semantic pathway during reading acquisition: the phonological pathway specializes in regular orthography-to-phonology mappings at the expense of exceptions, which are read by the semantic pathway. Damage to the semantic pathway causes "surface dyslexia" (where exceptions are selectively impaired), and damage to the phonological pathway causes "phonological dyslexia" (where nonwords are selectively impaired). According to this viewpoint, the syndrome of "deep dyslexia" (severe reading impairment, with meaning-based errors, such as reading the word *peach* as *apricot*) occurs when the phonological route is damaged and the semantic route is also partially impaired (which leads to semantic errors that are characteristic of the syndrome). Other highly successful connectionist models take the opposite line and directly implement both phonological routes, in line with standard views in cognitive neuropsychology (Coltheart et al., 1993). As well as exploring data on the breakdown of reading, there has also been a lively literature of (primarily) connectionist computational models of acquisition, although this issue is not explored here (Brown & Chater, 2003; Harm, McCandliss, & Seidenberg, 2003).

3.6. *Capturing the Psycholinguistic Data*

Moving from neuropsychological to experimental data, connectionist models of reading have been criticized for not modeling effects of specific lexical items (Spieler & Balota, 1997). One defense is that current models are too partial (e.g., containing no letter recognition and phonological output components) to model word-level effects (Seidenberg & Plaut, 1998). However, this challenge is taken up in a study in which an SRN is trained to pronounce words phoneme-by-phoneme (Plaut, 1999). The network can also refixate the input when unable to pronounce part of a word. The model performs well on words and nonwords, and fits empirical data on word length effects (Rastle & Coltheart, 1998; Weekes, 1997). Complementary work using a recurrent network focuses on providing a richer model of phonological knowledge and processing (Harm & Seidenberg, 1999, 2004), which may be importantly related to reading development (Bradley & Bryant, 1983).

Finally, it has been shown how a two-route model of reading might emerge naturally from a connectionist learning architecture (Zorzi et al., 1998). Direct links between orthographic input and phonological output learn to encode letter-to-phoneme correspondences (a phonological route) whereas links via hidden units spontaneously learn to handle exception words (a lexical route; Figure 17.2B). Here, as elsewhere, connectionist and indeed probabilistic models can provide persuasive instantiations of a range of theoretical positions.

3.7. *Probabilistic Approaches*

In recent work, there has been a trend toward developing probabilistic models of reading. One attraction of this approach is that it allows a clearer and more direct explanation of how the statistical structure of the orthography-phonology mapping and other factors such as word frequency lead to variations in reading performance. This approach can, to some degree, be viewed as providing a theoretical analysis of why some of the connectionist models work as they do. For example, many aspects of network behavior can be understood as depending on the regularity of the

orthography-phonology mapping at different levels of analysis (individual phonemes, trigrams, onsets/rimes, etc.); probabilistic models can provide a principled way of synthesizing regularities at different levels to produce predictions about how non-words should be read; dissonances between levels will suggest that processing is likely to be slowed (Brown, 1998). A comprehensive model by Norris (2006) provides examples of this approach. Moreover, probabilistic methods have also been extended recently to provide an "ideal" model of how eye movements should be controlled to maximize the expected information throughput to the reading system (Legge, Klitz, & Tjan, 1997).

4. Sentence Processing

Although symbolic models of sentence processing have been extensively developed in computational linguistics and many proposals concerning sentence processing have been framed in symbolic terms (e.g., Berwick & Weinberg, 1984; Crocker, 1996; Kurtzman, 1985; Yngve, 1960), much recent work oriented toward explaining specific psycholinguistic data has been carried out within the connectionist and probabilistic traditions.

Sentence processing provides a considerable challenge for connectionism. Some connectionists (Miyata, Smolensky, & Legendre, 1993) have built symbolic structures directly into the network, whereas others have chosen to construct a modular system of networks, each tailored to acquire different aspects of syntactic processing (Miikkulainen, 1996). However, the approach that has had the most impact involves directly training networks to discover syntactic structure from word sequences (Elman, 1991). This approach is the most radical approach, that is, it aims to dispense with traditional rule-based models of language and, indeed, any rigid distinction between grammar and processing, or competence and performance (Christiansen, 1992).

4.1. *Capturing Complexity Judgment and Reading Time Data*

One study has explored the learning of different types of recursion by training an SRN on small artificial languages (Christiansen & Chater, 1999). Christiansen and Chater reasoned that processing will be difficult to the extent that each piece of subsequent linguistic input is not predicted. They measured the average prediction error for the network, when trained on different sentence types, and predicted that errors should correlate with psycholinguistic data on processing difficulty. The results provided a good match with human data concerning the greater perceived difficulty associated with center-embedding in German compared with cross-serial dependencies in Dutch (Bach, Brown, & Marslen-Wilson, 1986). Moreover, error scores considered word by word from a related model were mapped directly onto reading times, providing an experience-based account for human data concerning the differential processing of singly center-embedded subject and object relative clauses by good and poor comprehenders (MacDonald & Christiansen, 2002).

Another approach to sentence processing involves a two-component model of ambiguity resolution, combining an SRN with a "gravitational" mechanism (Tabor, Juliano, & Tanenhaus, 1997). The SRN was trained in the usual way on sentences derived from a grammar. After training, SRN hidden unit representations for individual words were placed in the gravitational mechanism, and the latter was allowed to settle into a stable state. Settling times were then mapped onto word-reading times. The two-component model was able to fit data from several experiments concerning the interaction of lexical and structural constraints on the resolution of temporary syntactic ambiguities (i.e., garden path effects) in sentence comprehension. The two-component model has also been extended (Tabor & Tanenhaus, 1999) to account for empirical findings reflecting the influence of semantic role expectations on syntactic ambiguity

resolution in sentence processing (McRae, Spivey-Knowlton, & Tanenhaus, 1998).

4.2. *Capturing Grammaticality Rratings in Aphasia*

Some headway has also been made in accounting for data concerning the effects of acquired aphasia (i.e., language processing difficulties, typically resulting from damage to, or degeneration of, brain areas involved with language) on grammaticality judgments (Allen & Seidenberg, 1999). A bidirectional recurrent network (Figure 17.2C) was trained mutually to associate two input sequences: a sequence of word forms and a corresponding sequence of word meanings. The network was able to learn a small artificial language successfully, enabling it to regenerate word forms from meanings and vice versa. Grammaticality judgments were simulated by testing how well the network could recreate a given input sequence, allowing activation to flow from the provided input forms to meaning and then back again. Ungrammatical sentences were recreated less accurately than grammatical sentences; hence, the network was able to distinguish grammatical from ungrammatical sentences. The network was then "lesioned" by removing 10% of the weights in the network. Grammaticality judgments were then elicited from the impaired network for ten different sentence types from a classic study of aphasic grammaticality judgments (Linebarger, Schwartz, & Saffran 1983).The aphasic patients had problems with three of these sentence types, and the network fitted this pattern of performance impairment. Computational models of aphasia have also been formulated within the symbolic tradition (Haarmann, Just, & Carpenter 1997).

4.3. *Probabilistic Approaches to Sentence Processing*

In contrast to the connectionist models described earlier, probabilistic models have typically been viewed as complementary to symbolic linguistic representations, although many theorists take probabilistic methods to have substantial revisionist implications for traditional linguistic representations (e.g., Bod et al., 2003). Here, the focus is on how probabilistic ideas have led to a rethinking of structural accounts of parsing, such as minimal attachment (Frazier, 1979), as mentioned previously.

Structural principles have come under threat from psycholinguistic data that indicates that parsing preferences over structural ambiguities, such as prepositional phrase attachment, differ across languages, often in line with variations in observed corpus frequencies in these languages (e.g., Mitchell et al., 1995). Psycholinguists are increasingly exploring corpus statistics across languages, and parsing preferences seem to fit the probabilities evident in each language (Desmet et al., in press; Desmet & Gibson, 2003).

Structural parsing principles also have difficulty capturing the probabilistic influence of lexical information. Thus, a structural principle finds it difficult to account for the difference in parsing preference between *the astronomer saw the planet with a telescope* and *the astronomer saw the star with a moon*. The probabilistic approach seems useful here because it seems important to integrate the constraint that seeing-with-telescopes is much more likely than seeing-with-moons.

One way to capture these constraints aims to capture statistical (or even rigid) regularities between words. For example, "lexicalized" grammars, which carry information about what material co-occurs with specific words, substantially improve computational parsing performance (Charniak, 1997; Collins, 2003). More generally, the view that parsing preferences are determined by the integration of many "soft" constraints, rather than by any single principle, structural or otherwise, is compatible with both connectionist and probabilistic frameworks (Seidenberg & MacDonald, 1999).

4.4. *Plausibility and Statistics*

Statistical constraints between words are, however, a crude approximation of what

sentences are *plausible*. In off-line judgement tasks, for example, where people assign explicit ratings of plausibility, people can use world knowledge, the understanding of the social and environmental context, pragmatic principles, and much more, to determine what people might plausibly say or mean. Determining whether a statement is plausible may involve determining how likely it is to be true, but also whether, given the present context, it might plausibly be *said*. The first issue requires a probabilistic model of general knowledge (Pearl, 1988). The second issue requires engaging "theory of mind" (inferring the other's mental states) and invoking principles of pragmatics. Models of these processes, probabilistic or otherwise, are very preliminary (Jurafsky, 2003).

A fundamental theoretical debate concerns whether plausibility is used on-line in parsing decisions. Are statistical dependencies between words used as a computationally cheap surrogate for plausibility? Or are both statistics and plausibility deployed on-line, perhaps in separate mechanisms? Eye-tracking paradigms (Tanenhaus et al., 1995; McDonald & Shillcock, 2003) have been used to suggest that both factors are used on-line, although this interpretations is controversial. However, recent work indicates that probabilistic grammar models often predict the time course of processing (Hale, 2003; Jurafsky, 1996; Narayanan & Jurafsky, 2002).

4.5. *Is the Most Likely Parse Favored?*

In the probabilistic framework, it is typically assumed that on-line ambiguity resolution favors the most probable parse. Yet, Chater, Crocker, and Pickering (1998) suggest that, for a serial parser, whose chance of "recovery" is highest if the "mistake" is discovered soon, this is an oversimplification. In particular, they suggest that because parsing decisions are made *on-line* (Pickering, Traxler, & Crocker, 2000), there should be a bias to choose interpretations that make *specific* predictions, which might rapidly be falsified. For example, in the phrase *John realized his . . .*, the more probable interpretation is that *realized* introduces a sentential complement (i.e., *John realized [that] his . . .*). On this interpretation, the rest of the noun phrase after *his* is unconstrained. By contrast, the less probable transitive reading (*John realized his goals/potential/objectives*) places very strong constraints on the subsequent noun phrase. Perhaps, then, the parser should favor the more specific reading because if wrong, it may rapidly and successfully be corrected. Chater, Pickering, and Crocker (1998) provide a Bayesian analysis of "optimal ambiguity resolution" capturing such cases. The empirical issue of whether the human parser follows this analysis (Pickering et al., 2000) is not fully resolved. Note, too, that parsing preferences appear to be influenced by additional factors, including the linear distance between the incoming word and the prior words to which it has a dependency relation (Grodner & Gibson, 2005).

Overall, connectionist and probabilistic computational proposals have allowed a more fine-grained match with psycholinguistic data than obtained by early symbolic models. The question of how far models of sentence processing, considering the full complexity of natural language syntax and the subtlety of compositional semantics, can avoid adopting traditional symbolic representations, as postulated by linguistic theory, remains controversial.

5. Language Acquisition

Chomsky (1965) frames the problem of language acquisition as follows: The child has a hypothesis-space of candidate grammars and must choose, on the basis of (primarily linguistic) experience, one of these grammars. From a probabilistic standpoint, each candidate grammar is associated with a prior probability, and these probabilities will be modified by experience using Bayes' theorem (see Chapter 3 in this volume). The learner will presumably choose a language with high, and perhaps the highest, posterior probability.

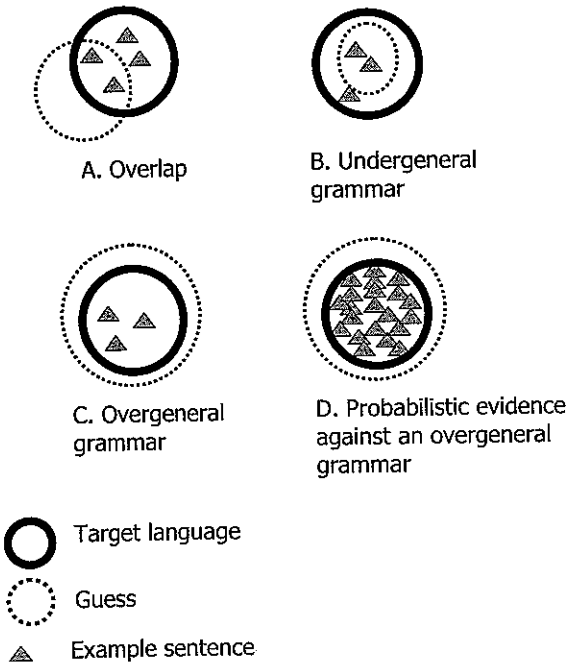


Figure 17.3. The problem of recovery from overgeneral grammars. Suppose that sentences (gray triangles) are generated according to a true grammar, indicated by the unbroken circles. The learner considers an alternative, incorrect, grammar, shown by the broken circles. How can the learner realize that its current guess is incorrect and that it needs to search for an alternative grammar? (A) When the grammars partially overlap, a sentence will eventually be encountered that is outside the hypothesized grammar. (B) The same is true when the learner's proposed grammar is undergeneral, that is, the sentences of the hypothesized grammar are all allowed in the true language, but the proposed grammar does not allow some sentences that are grammatical. (C) A problem arises when the learner's grammar is overgeneral because all the sentences that are encountered by the learner fit the overgeneral language; hence, there is no decisive way of falsifying the overgeneral language from observed sentences. Of course, if the learner *produces* an illegitimate sentence, it may obtain feedback that this sentence is not acceptable, but it is widely, although controversially, argued that such feedback is not required for language acquisition. The puzzle of how to recover from postulating an overgeneral grammar, which arises in a range of guises, has been seen as so serious as to pose a "logical" problem for language acquisition (e.g., Baker & McCarthy, 1981; see MacWhinney, 2004, for discussion). (D) If, however, the learner finds that only a portion of the space of possible sentences is actually used, for example,

5.1. *The Poverty of the Stimulus?*

Chomsky (1968, 1980) influentially argued, as noted earlier, that the learning problem is unsolvable without strong prior constraints on the language, given the "poverty" (i.e., partiality and errorfulness) of the linguistic stimulus. Indeed, Chomsky (1981) argued that almost all syntactic structure, aside from a finite number of binary parameters, must be innate. Independent mathematical results by Gold (1967) indicated that, under certain assumptions, learners provably cannot converge on a language even "in the limit" as the corpus becomes indefinitely large (for discussion, see MacWhinney, 2004; Rohde & Plaut, 1999). In essence, the problem is that the learner seems to have no way of guaranteeing recovery from formulating an overgeneral grammar, at least if it is restricted to observing sentences in the language. This is because all the sentences that it hears are allowed by the overgeneral grammar, and hence the learner appears to have no impetus to switch to a new grammar (Figure 17.3).

A probabilistic standpoint yields more positive learnability results. For example, Horning (1969) proved that phrase structure grammars are learnable (with high probability) to within a statistical tolerance if sentences are sampled as independent, identically distributed data. Chater and Vitányi (2007; Chater, 2004, gives a brief summary) generalize to a language that is generated by any computable process (i.e., sentences may be interdependent,

Figure 17.3 (*cont.*)
those that fit with some other grammar, then the learner should become increasingly persuaded that the grammar is overgeneral. This argument, although intuitively appealing, is not straightforward, however, because, of course, the learner's experience of the language is always a finite subset of an infinite set of possible sentences. A key observation is that an alternative *simple* and less general grammar is available that captures all observed sentences. This type of argument can be made rigorous (e.g., Horning, 1969; Chater, 2004).

as they are in real language corpora, and sentences may be generated by any computable process, i.e., the highest level in the Chomsky hierarchy). They show that prediction, grammaticality, and semantics are all learnable to a statistical tolerance. These results are "ideal"; however, they consider what would be learned if the learner could find the shortest representation of linguistic data and use this representation as the basis for prediction, grammaticality judgements, and so on. In practice, the learner may find a short code, but not the shortest, and theoretical results are not available for this case. Nonetheless, from a probabilistic standpoint, learning looks less intractable, partly because learning need only succeed with high probability and to an approximation (speakers may learn slightly different idiolects).

5.2. *Computational Models of Language Learning*

Yet, the question of learnability and the potential need for innate constraints remain. Machine learning methods have successfully learned small artificial context-free languages (e.g., Anderson, 1977; Lari & Young, 1990), but profound difficulties in extending these results to real language corpora have led computational linguists to focus on learning from parsed trees (Charniak, 1997; Collins, 2003), presumably not available to the child. Connectionist models are restricted to small artificial languages (Elman, 1990; Christiansen & Chater, 1999) and, despite having considerable psychological interest, they often do not scale well (though see Reali, Christiansen, & Monaghan, 2003).

Klein and Manning (2002, 2004) have recently made substantial steps toward solving the problem of deriving syntactic constituency from a corpus of unlabelled, unparsed text. Klein and Manning (2002) extended the success of distributional clustering methods for learning word classes (Redington et al., 1998; Schütze 1998), discussed later. Roughly, they classify the categories of phrases by grouping together phrases that have similar contexts (context here concerns the word immediately preceding and immediately following the phrase). As discussed later, this corresponds to a statistical version of the distributional test in linguistics. Klein and Manning (2004) combine this work with a system for learning linguistic dependency relations. The dependency model uses data on which words occur together, with two additional and crucial constraints: that dependencies between nearby words are preferred and a preference for words to have few dependencies. Klein and Manning's work shows that central features of language, phrase structure and dependency relations can be learned to a good approximation from unlabelled language – clearly a task crucial to child language acquisition.

This work is a promising demonstration of empirical language learning from a probabilistic standpoint, but most linguistic theories use richer structures than surface phrase structure trees. Moreover, learning the syntactic regularities of language should, presumably, be in the service of learning how to map linguistic forms to meanings. In the probabilistic tradition, there is some work on mapping to meaning representations of simple data sets (Zettlemoyer & Collins, 2005) and work on unsupervised learning of a mapping from surface text to semantic role representations (Swier & Stevenson, 2005). There is also a related tradition of work, especially on thematic role assignment, in the connectionist tradition (Lupyan & Christiansen, 2002; McClelland & Kawamoto, 1986; St. John, 1992).

5.3. *Poverty of the Stimulus, Again*

The status of Chomsky's (1965) poverty of the stimulus argument remains unclear, beginning with the question of whether children really do face a poverty of linguistic data (see the debate between Pullum & Scholz, 2002, and Legate & Yang, 2002). Perhaps no large and complex grammar can be learned from the child's input, or perhaps certain specific linguistic patterns (e.g., perhaps encoded in an innate universal

grammar) are in principle unlearnable. Interestingly, however, Reali and Christiansen (2005) have shown that both probabilistic and connectionist methods can successfully be applied to learning an apparently problematic linguistic construction, auxiliary fronting, suggesting that more linguistic phenomena may be learnable than is typically assumed in linguistics (e.g., Chomsky, 1957).

Presently, theorists using probabilistic methods diverge widely on the severity of the prior "innate" constraints they assume. Some theorists focus on applying probability to learning parameters of Chomskyan Universal Grammar (Gibson & Wexler, 1994; Niyogi, 2006); others focus on learning relatively simple aspects of language, such as learning morphological structure, or learning approximate syntactic or semantic categories, with relatively weak prior assumptions.

5.4. *Acquiring Morphological Structure*

A key issue in computational models of morphological processing and acquisition is how computational analysis has addressed a key theoretical question: whether inflectional morphology requires two "routes," one to handle regular morphology (e.g., "go" → "went") or whether a single computational mechanism can account for both rules and exceptions paralleling the single route vs. dual route debate in reading, discussed previously. Studies with idealized languages patterned on English past tense morphology suggest that a single route may handle both cases (Hahn & Nakisa, 2000). However, Prasada and Pinker (1993) argued that the success of these models results from the distributional statistics of English. Many regular English /-ed/ verbs have low token frequencies, which a connectionist model can handle by learning to add /-ed/ as a default. For irregular verbs, token frequency is typically high, allowing the network to override the default. Prasada and Pinker argued that a default regular mapping with both low type and token frequency could not be learned by a connectionist network. The putative default /-s/ inflection of plural nouns in German appears to provide an example of such a "minority default mapping." Marcus et al. (1995) proposed that the German plural system must be modeled by two routes: a pattern associator, which memorizes specific cases (both irregular and regular), and a default rule (add /s/), which applies when the connectionist pattern associator fails.

Hahn and Nakisa (2000) asked whether single route associative models (they tested two exemplar-based learning models and a simple feed-forward connectionist net with one hidden layer) could learn the German plural system and generalize appropriately to novel regular and irregular nouns. Their models' task was to predict to which of 15 different plural types the input stem belonged. The inputs to the learning mechanisms were phonetic representations of 4,000 German nouns taken from the CELEX database (token frequency was ignored). All models showed good performance in predicting the plural form of 4,000 unseen nouns, and the connectionist model obtained the best performance, at over 80% correct.

Crucially, Hahn and Nakisa (2000) also simulated the Marcus et al. (1995) model by assuming that any test word which is not close to a training word, according to the associative model (for which the lexical memory fails), will be dealt with by a default "add /-s/" rule. The associative models were trained on the irregular nouns, and the models were tested as before. They found that for all three models, the presence of the rule led to a decrement in performance. In general, the higher the threshold for memory failure (the more similar a test item had to be to a training item to be irregularized via the associative memory), the greater the decrement in performance. The use of a default rule could only have improved performance for regular nouns occupying regions of phonemic space surrounding clusters of irregulars. Hahn and Nakisa's findings demonstrate that very few regular nouns occur in these regions in the German lexicon. The extension of Hahn and Nakisa's findings

to the production of the plural form (instead of merely indicating the plural type) and to more realistic input (for instance, taking account of token frequency) remains to be performed. Further work might also focus on the extent to which different single- and dual-route models are able to capture changes in detailed error patterns of under-regularization- and overregularization during development. Another interesting topic for future work is the processing of derivational, rather than inflectional, morphology (e.g., Plaut & Gonnerman, 2000).

5.5. *Acquiring Syntactic Categories*

The problem of categorizing phrases using distributional methods from unlabelled text (Klein & Manning, 2002) has been discussed. A more basic question is how does the child acquire lexical syntactic categories, such as noun and verb. This problem encompasses both discovering that there are different classes and ascertaining which words belong to each class. Even for theorists who assume that the child innately possesses a universal grammar and syntactic categories (as is assumed in the traditional Chomskyan framework), identifying the category of particular words must primarily be a matter of learning. Universal grammatical features can only be mapped on to the specific surface appearance of a particular natural language once the identification of words with syntactic categories has been made, although once some identifications have been made, it may be possible to use prior grammatical knowledge to facilitate further identifications. The contribution of innate knowledge to initial linguistic categories must be relatively slight. Both language-external and language-internal cues may be relevant to learning syntactic categories. One language-external approach, semantic bootstrapping, exploits the putative correlation between linguistic categories (in particular, noun and verb) and the child's perception of the environment (in terms of objects and actions). This may provide a means of "breaking in" to the system of syntactic categories. Also, there may be many relevant language-internal factors: regularities between phonology, prosody and distributional analysis, both over morphological variations between lexical items (e.g., affixes such as "-ed" are correlated with syntactic category; Maratsos & Chalkley, 1980; see also Onnis & Christiansen, 2005), and at the word level.

Here, the focus is on this last approach, which has a long history, although this method of finding word classes has often been dismissed on a priori grounds within the language learning literature. The "distributional test" in linguistics is based on the observation that if all occurrences of word A can be replaced by word B without loss of syntactic acceptability, then they share the same syntactic category. For example, dog can be substituted freely for cat in phrases such as: *the cat sat on the mat, nine out of ten cats prefer . . .*, indicating that these items have the same category. The distributional test is not a foolproof method of grouping words by their syntactic category, because distribution is a function of many factors other than syntactic category (such as word meaning). Thus, for example, *cat* and *barnacle* might appear in very different contexts in some corpora, although they have the same word class. Nevertheless, it may be possible to exploit the general principle underlying the distributional test to obtain useful information about word classes. One approach is to record the contexts in which the words to be classified appear in a corpus of language and group together words with similar distributions of contexts. Here, context is defined in terms of co-occurrence statistics.

Redington et al. (1998) used a window of two words before and after each target word as context. Vectors representing the co-occurrence statistics for these positions were constructed from a 2.5 million-word corpus of transcribed adult speech taken from the CHILDES corpus (MacWhinney & Snow, 1985), much of which was child-directed). The vectors for each position were concatenated to form a single vector for each of 1,000 target words. The similarity of distribution between the vectors was calculated using Spearman's rank correlation, and

hierarchical cluster analysis was used to group similar words together.

This approach does not partition words into distinct groups corresponding to the syntactic categories, but produces a hierarchical tree, or dendrogram, whose structure reflects to some extent the syntactic relationships between words. Figure 17.4A shows the high-level structure of the dendrogram resulting from the previous analysis. Figure 17.4B shows part of the Adjective cluster in Figure 17.4A, illustrating how statistical distributional analysis reflects syntactic and semantic information at a very fine level.

A quantitative analysis of the mutual information between the structure of the dendrogram and a canonical syntactic classification of the target words (defined as their most common syntactic usage in English) as a percentage of the joint information in both the derived and canonical classifications revealed that at all levels of similarity, the dendrogram conveyed useful information about the syntactic structure of English. Words that were clustered together tended to belong to the same syntactic category, and words that were clustered apart tended to belong to different syntactic categories. Thus, computational analysis of real language corpora shows that distributional information at the word level is highly informative about syntactic category, despite a priori objections to its utility (see Monaghan, Chater, & Christiansen, 2005). Similar results, typically on a smaller scale, have been obtained from hidden-unit analysis of connectionist networks (e.g., Elman, 1990; although such results also arise when the network is untrained; Kolen, 1994).

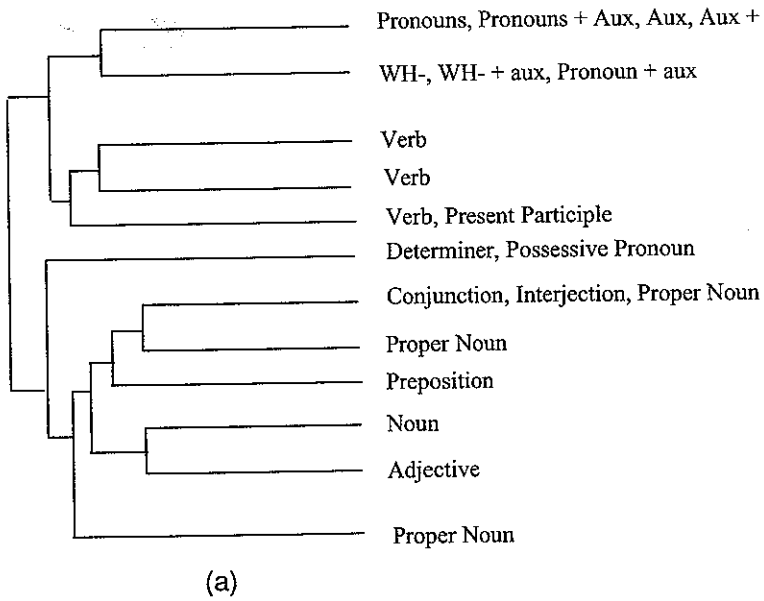
5.6. *Acquiring Lexical Semantics*

Acquiring lexical semantics involves identifying the meanings of particular words. Even for concrete nouns, this problem is complicated by the difficulty of detecting which part of the physical environment a speaker is referring to. Even if this can be ascertained, it may still remain unclear whether the term used by the speaker refers to a particular object, a part of that object, or a class of objects. For abstract nouns and other words that have no concrete referents, these difficulties are compounded further.

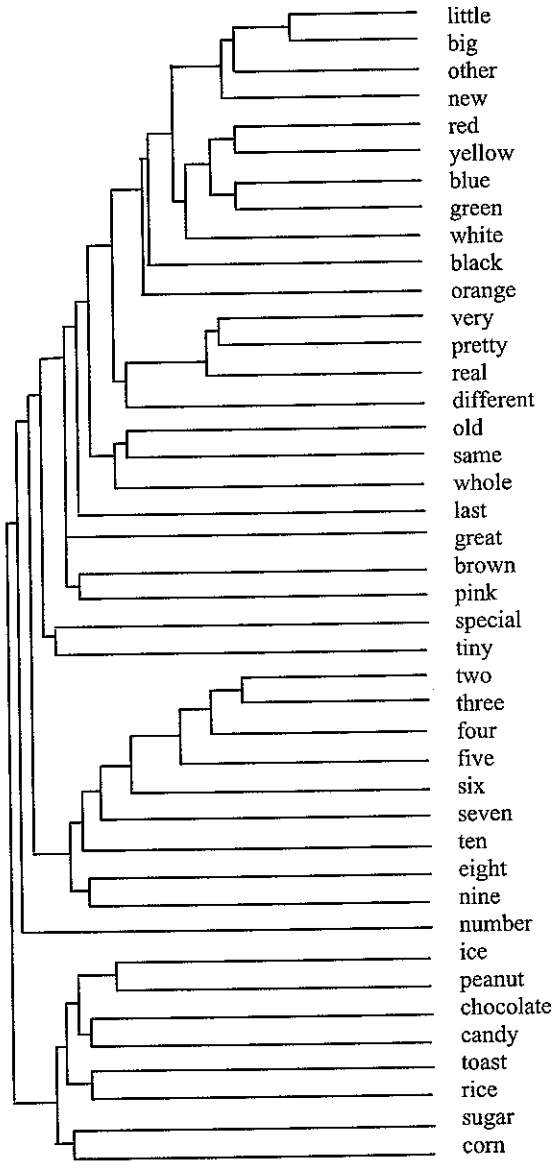
Presumably, the primary sources of information for the development of lexical semantics are language-external. Relationships between the child and the physical environment, and especially the social environment, are likely to play a major role in the development of lexical semantic knowledge. However, it also seems plausible that language-internal information might be used to constrain the identification of the possible meaning of words. For instance, just as semantics might constrain the identity of a word's syntractic category (words referring to concrete objects are likely to be nouns), knowing a word's syntactic category provides some constraint on its meaning; in general, knowing that a word is a noun, perhaps because it occurs in a particular set of local contexts, implies that it will refer to a concrete object or an abstract concept, rather than an action or process.

Because there are potentially informative relationships between aspects of language at all levels, this means that even relatively low-level properties of language, such as morphology and phonology, might provide some constraints on lexical semantics. Gleitman has proposed that syntax is a potentially powerful cue for the acquisition of meaning. Gleitman assumes that the child possesses a relatively high degree of syntactic knowledge. However, an examination of Figure 17.4B shows that the distributional method used earlier to provide information about syntactic categories also captures some degree of semantic relatedness without any knowledge of syntax proper. More direct methods for deriving semantic relationships have been proposed (e.g., Landauer & Dumais, 1997; Lund & Burgess, 1996; Schütze, 1993).

These statistical approaches do, however, have a somewhat arbitrary quality. Griffiths and Steyvers (2004) have more recently developed a more rigorous Bayesian approach, in which the words in a text are viewed as generated from a mixture of "topics,"



(a)



(b)

Figure 17.4 Word clusters from distributional information, based on Redington, Chater, and Finch (1998). Analysis was conducted over a large corpus of transcribed adult speech from the CHILDES database (MacWhinney & Snow, 1985). (A) The overall structure of a cluster of the lexicon in which the syntactic labels, added by hand, classify the category of the vast majority of the words correctly compared with standard classifications. The numbers in parentheses indicate the number of lexical items in each category. (B) A close-up of part of the adjective cluster. Note the fine-grained semantic groupings obtained, such as groupings of color and number words. (Reprinted with permission of the Cognitive Science Society, from Redington, M., Chater, N., & Finch, S. [1998]. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 425–469.)

TRANSPORTATION	BEANS	ROOM	BARN	CREEK	SAVINGS
CARS	POTATOES	HOUSE	CHICKENS	BANK	MONEY
TRUCKS	POTATO	BED	HOUSE	STREAM	ACCOUNT
ROADS	TOMATOES	TABLE	HEN	SIDE	INTEREST
TRAVEL	SWEET	KITCHEN	BIG	WOODS	ACCOUNTS
TRAINS	VEGETABLES	ROOMS	FARM	FEET	FUNDS
AIRPLANES	CORN	BEDROOM	COWS	MEADOW	LOAN
AUTOMOBILES	LETTUCE	DOOR	HAY	DEEP	BANK
HIGHWAYS	CARROTS	WALLS	STRAW	RIVER	DEPOSITS
CARRY	BEAN	FLOOR	LANTERN	BUSHES	MUTUAL
TRANSPORT	EAT	CHAIR	HENS	RAN	DEPOSIT
GOODS	SQUASH	WALL	HORSE	BROOK	CHECKING
BUSES	CABBAGE	LIVING	STALL	STOOD	HIGHER
BUILT	TOMATO	WINDOWS	SHED	POOL	INSTITUTION
CITIES	PEAS	HALL	PIGS	WATER	OFFER
MOVE	BANANAS	DINING	ROOSTER	EDGE	FUND
FREIGHT	PEPPERS	FURNITURE	COW	BRANCHES	ASSOCIATIONS
RAILROADS	CABBAGES	CHAIRS	HORSES	TREES	PROVIDE
PASSENGERS	SPROUTS	CURTAINS	CHICKEN	LAY	INVEST
SHIPS	PEANUTS	SHELVES	STALLS	WALKED	EARN

Figure 17.5. Semantic relations between lexical items, learned from distributional information. Six semantic "topics" derived from a large text corpus (the TASA corpus), using the method of Griffiths & Steyvers (2004) and chosen from 1,700 topics used in this analysis. The top twenty most frequent words for each topic are shown in rank order. They correspond to transport, food, furniture, barnyard animals, pastoral, and finance topics. Note that *bank* occurs in both the latter contexts, indicating that multiple readings of a word can be recognized.

and the topics themselves are inferred from the data. This Bayesian approach provides an elegant way of finding semantic categories from text. Thus, good approximations to syntactic categories and semantic classes have been learned by clustering items based on their linear distributional contexts (e.g., the distribution over the word that precedes and follows each token of a type) or broad topical contexts (see Figure 17.5). One can even simultaneously cluster words exploiting local syntactic and topical similarity (Griffiths et al., 2005).

Grouping words that are semantically related is only a small part of the problem of learning lexical semantics, of course. One particularly pressing problem is that such analyses merely relate words to each other, rather than connecting them to objects in the world. The problem of relating words to the referents (e.g., as presented in perceptual input for words with concrete referents) raises very large computational challenges. Nonetheless, some interesting work has been carried out that begins to address this problem (e.g., Regier, 2005; Roy, 2005).

6. Conclusion and Future Directions

Linguistics has traditionally viewed language as a symbolic system, governed by a rich system of rules; yet, computational models of human language processing have focused on graded and probabilistic aspects of language structure and processing. Connectionist and probabilistic computational accounts of psycholinguistic phenomena have been proposed, ranging from speech processing to phonology, morphology, reading, syntax, semantics, and language production. Moreover, as has been seen, connectionist and probabilistic approaches have provided both new theoretical perspectives and specific computational models of a range of aspects of language acquisition, typically emphasizing the importance of information in the linguistic input far more than the strongly nativist tradition in Chomskyan linguistics.

There is reason to expect, nonetheless, that future developments in computational modeling of psycholinguistic phenomena will involve an interplay between all three perspectives on language. Language has a

substantial symbolic component, even if it also has much graded and probabilistic structure; a wide variety of recent work in linguistics, much of it inspired by or directly rooted in computational modeling, proposes that symbolic and probabilistic aspects of language must be explained simultaneously (e.g., Goldberg, 2006). There is a variety of overlapping ways in which rule-based and probabilistic factors may interact: The set of potentially conflicting linguistic rules to apply may be determined using probabilistic methods (e.g., Smolensky & Legendre, 2006); rules may be embodied directly in stochastic grammars (e.g., Charniak, 1997); rules, and perhaps also their exceptions, may be probabilistically approximated using connectionist networks (Christiansen & Chater, 2001). The project of building deeper models of human language processing and acquisition involves paying attention to both rules and to graded/probabilistic structure in language. At the same time, the project of computational modeling must be in close dialogue with both theoretical work in linguistics and, perhaps most crucially, with the increasingly sophisticated and detailed body of empirical data on how people use and acquire language.

Acknowledgments

We would like to thank Tom Griffiths, Chris Manning, and Martin Redington for input into this work and two anonymous reviewers for their comments on an earlier draft of this chapter. Nick Chater was partially supported by ESRC grant RES-000-22-1120 and a Leverhulme Trust Major Research Grant. Morten Christiansen was supported by a Charles A. Ryskamp Research Fellowship from the American Council of Learned Societies.

References

Aho, A. V., & Ullman, J. D. (1972). *The theory of parsing, translation and compiling* (Vol. 1). Englewood Cliffs, NJ: Prentice Hall.

Allen, J., & Seidenberg, M. S. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *The emergence of language* (pp. 115–151). Mahwah, NJ: Lawrence Erlbaum.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Anderson, J. A. (1977). Induction of augmented transition networks. *Cognitive Science, 1*, 125–157.

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.

Baker, C. L., & McCarthy, J. J. (Eds.). (1981). *The logical problem of language acquisition*. Cambridge, MA: MIT Press.

Bates, E. A., & Elman, J. L. (1993). Connectionism and the study of change. In M. H. Johnson (Ed.), *Brain and cognitive development: A reader* (pp. 623–642). Oxford, UK: Blackwell.

Berwick, R., (1986). *The acquisition of linguistic knowledge*. Cambridge, MA: MIT Press.

Berwick, R., & Weinberg, A. (1984). *The grammatical basis of linguistic performance*. Cambridge, MA: MIT Press.

Besner, D., Twilley, L., McCann, R. S., & Seergobin, K. (1990). On the connection between connectionism and data: are a few words necessary? *Psychological Review, 97*, 432–446.

Bod, R. (1998). *Beyond grammar: An experience-based theory theory of language*. Stanford, CA: CSLI Publications.

Bod, R., Hay, J., & Jannedy, S. (Eds.). (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Bradley, L., & Bryant, P. (1983). Categorizing sounds and learning to read: a causal connection. *Nature, 301*, 419–421.

Brown, G. D. A. (1998). The endpoint of reading instruction: The ROAR model. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning Literacy* (pp. 121–138). Mahwah, NJ: Erlbaun.

Brown, G. D. A., & Chater, N. (2003). Connectionist models of children's reading. In T. Nunes & P. E. Bryant (Eds.), *Handbook of children's literacy* (pp. 67–89). Dordrecht: Kluwer.

Cairns, P., Shillcock, R. C., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A

bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33,* 111–153.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In B. F. Kuipers & B. Webber (Eds.), *Proceedings of the 14th National Conference on Artificial Intelligence* (pp. 598–603). Cambridge, MA: AAAI Press.

Chater, N. (2004). What can be learned from positive evidence?*Journal of Child Language, 31,* 915–918.

Chater, N., Crocker, M. W., & Pickering, M. J. (1998). The rational analysis of inquiry: The case of parsing. In M. Oaksford, & N. Chater (Eds.), *Rational models of cognition* (pp. 441–469). Oxford, UK: Oxford University Press.

Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10,* 335–344.

Chater, N., & Oaksford, M. (1990). Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn. *Cognition, 34,* 93–107.

Chater, N., Tenenbaum, J. B., & Yuille, A. (Eds.) (2006). Probabilistic models of cognition [Special issue]. *Trends in Cognitive Sciences, 10.*

Chater, N., & Vitányi, P. M. B. (2007). 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology, 51,* 135–163.

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

Chomsky, N. (1959). A review of B.F. Skinner's verbal behavior. *Language, 35,* 26–58.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1968). *Language and mind.* New York: Harcourt Brace Jovanovich.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences, 3,* 1–61.

Chomsky, N. (1981). *Lectures on government and binding.* Dordrecht: Foris.

Chomsky, N., & Schützenberger, M. P. (1963). The algebraic theory of context free languages. In P. Braffort & D. Hirschberg (Eds.), *Computer programming and formal languages* (pp. 118–161). Amsterdam: North-Holland.

Christiansen, M. H. (1992). The (non) necessity of recursion in natural language processing. In J. Kolodner & C. Riesbeck (Eds.), *Proceedings of the 14th Annual Cognitive Science Society Conference* (pp. 665–670). Hillsdale, NJ: Lawrence Erlbaum.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: a connectionist model. *Language and Cognitive Processes, 13,* 221–268.

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

Christiansen, M. H., & Chater, N. (Eds.). (2001). *Connectionist psycholinguistics.* Westport, CT: Ablex.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational Linguistics, 29*(4), 589–637.

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.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256.

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 parser. In D. R. Dowty, L. Kartunnen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 320–358). New York: Cambridge University Press.

Crocker, M. W. (1996). *Computational psycholinguistics: An interdisciplinary approach to the study of language.* Boston: Kluwer.

Daelemans, W., & Van den Bosch, A. (2005). *Memory-based language processing.* Cambridge, UK: Cambridge University Press.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modelling of neural systems.* Cambridge, MA: MIT Press.

Desmet, T., De Baecke, C., Drieghe, D., Brysbaert, M., & Vonk, W. (in press). Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language and Cognitive Processes.*

Desmet, T., & Gibson, E. (2003). Disambiguation preferences and corpus frequencies in noun phrase conjunction.*Journal of Memory and Language, 49,* 353–374.

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

Elman, J. L. (1991). Distributed representation, simple recurrent networks, and grammatical structure. *Machine Learning, 7,* 195–225.

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

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

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27,* 143–165.

Fanselow, G., Féry, C., Vogel, R., & Schlesewsky, M. (Eds.). (in press). *Gradience in grammar: Generative perspectives.* Oxford, UK: Oxford University Press.

Ferreira, F., & Clifton, C. J. R. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25,* 348–368.

Fodor, J. A. (1983). *Modularity of mind.* Cambridge, MA: MIT Press.

Fodor, J. A., Bever, T. G., & Garrett, M. F. (1974). *The psychology of language: An introduction to psycholinguistics and generative grammar.* New York: McGraw-Hill.

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

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 257–287). Amsterdam: North-Holland.

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies.* Bloomington: Indiana University Linguistics Club.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition, 13,* 187–222.

Ganong, W. F. I. (1980). Phonetic categorisation in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 110–125.

Gaskell, M. G., & Marslen-Wilson, W. D. (1995). Modeling the perception of spoken words. In *Proceedings of the 17th Annual Cognitive Science Conference* (pp. 19–24). Hillsdale, NJ: Lawrence Erlbaum.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997a). Discriminating local and distributed models of competition in spoken word recognition. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Cognitive Science Conference* (pp. 247–252). Hillsdale, NJ: Lawrence Erlbaum.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997b). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12,* 613–656.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science, 23,* 439–462

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

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry, 25,* 407–454.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10,* 447–474.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language.* Oxford, UK: Oxford University Press.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl. 1), 5228–5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In B. Schölkopf, J. Platt & T. Hofmann (Eds.), *Advances in Neural Information Processing Systems, 17.*

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science, 29,* 261–291.

Haarmann, H. J., Just, M. A., & Carpenter, P. A. (1997). Aphasic sentence comprehension as a resource deficit: A computational approach. *Brain and Language, 59,* 76–120.

Hahn, U., & Nakisa, R. C. (2000). German inflection: Single or dual route? *Cognitive Psychology, 41,* 313–360.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research, 32,* 101–123.

Harm, M., McCandliss, B. D., & Seidenberg, M. S. (2003). Modeling the successes and failures of interventions for disabled readers. *Scientific Studies of Reading, 7,* 155–182.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106,* 491–528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading:

Division of labor between visual and phonological processes. *Psychological Review, 111,* 662–720.

Hay, J., & Baayen, H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences, 9,* 342–348.

Hirschfeld, L. A., & Gelman, S. A. (Eds). (1994). *Mapping the mind: Domain specificity in cognition and Culture.* New York: Cambridge University Press.

Hockema, S. A. (in press). Finding words in speech: An investigation of American English. *Language Learning and Development.*

Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2000). *Introduction to automata theory, languages and computability.* Boston: Addison-Wesley.

Horning, J. J. (1969). *A study of grammatical inference* (Tech. Rep. CS 139). Stanford, CA: Computer Science Department, Stanford University.

Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition, *Technometrics, 33,* 251–272.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20,* 137–194.

Jurafsky, D. (2003). Pragmatics and computational linguistics. In L. R. Horn & G. Ward (Eds.), *Handbook of pragmatics.* Oxford, UK: Blackwell.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language, 32,* 474–516.

Klein, D., & Manning, C. (2002). A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the ACL, 2004* (pp. 128–135).

Klein, D., &. Manning, C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the ACL.*

Kolen, J. F. (1994). The origin of clusters in recurrent network state space. In A. Ram & K. Eiselt (Eds.), *The Proceedings Sixteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51,* 141–178.

Kurtzman, H. (1985). *Studies in syntactic ambiguity resolution.* Unpublished doctoral dissertation, MIT, Cambridge, MA.

Labov, W. (1972). *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

Lari, K., & Young., S. Y. (1990). The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language, 4,* 35–56.

Legate, J. A., & Yang, C. D. (2002). Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review, 19,* 151–162.

Legge G. E., Klitz T. S., & Tjan B. S. (1997). Mr. Chips: An ideal observer model of reading. *Psychological Review, 104,* 524–553.

Lightfoot, D. (1991). *How to set parameters: Arguments from language change.* Cambridge, MA: MIT Press.

Linebarger, M. C., Schwartz, M. F., & Saffran, E. M. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition, 13,* 361–392.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers, 28,* 203–208.

Lupyan, G., & Christiansen, M. H. (2002). Case, word order, connectionist modeling. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 596–601). Mahwah, NJ: Erlbaum.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: A comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109,* 35–54.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101,* 676–703.

MacWhinney, B. (2004). Multiple solutions to the logical problem of language acquisition. *Journal of Child Language, 31,* 883–914.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12,* 271–296.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69,* 548–558.

Maratsos, M., & Chalkley, M. (1980). The internal language of children's syntax: The

ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's language* (Vol. 2). (pp. 127–214). New York: Gardner Press.

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

Marcus, G. F. (2000). Children's overregularization and its implications for cognition. In P. Broeder & J. Murre (Eds.), *Models of language acquisition: Inductive and deductive approaches* (pp. 154–176). New York: Oxford University Press.

Marcus. G. F., Brinkman, U., Clahsen, H., Weise, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology, 29*, 189–256.

Marcus, M. P. (1980). *A theory of syntactic recognition for natural language*. Cambridge, MA: MIT Press.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101*, 653–675.

Marslen-Wilson, W., & Welsh, A. (1978). Processing instructions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29–63.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.

McClelland, J. L., & Kawamoto, A. H. (1986). Mechanisms of sentence processing: Assigning roles to constituents. In J. L. McClelland, & D. E. Rumelhart (Eds.), *Parallel distributed processing* (pp. 272–325). Cambridge, MA: MIT Press.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities. *Psychological Science, 14*, 648–652.

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

Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science, 20*, 47–73.

Miller, G. A. (1962). Some psychological studies of grammar. *American Psychologist, 17*, 748–762.

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research, 24*, 469–488.

Miyata, Y., Smolensky, P., & Legendre, G. (1993). Distributed representation and parallel distributed processing of recursive structures. In W. Kintsch (Ed.), *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society* (pp. 759–764). Hillsdale, NJ: Lawrence Erlbaum.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential contribution of phonological and distributional cues in grammatical categorisation. *Cognition, 96*, 143–182.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76*, 165–178.

Narayanan, S., & Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading time in sentence processing. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 59–65), Cambridge, MA: MIT Press.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Norris, D. (1993). Bottom-up connectionist models of "interaction." In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing* (pp. 211–234), Hillsdale, NJ: Lawrence Erlbaum.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*, 189–234.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327–357.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*, 299–370.

Onnis, L., & Christiansen, M. (2005). New beginnings and happy endings: Psychological plausibility in computational models of language acquisition. In B. G. Bara, L. Barsalov & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufman.

Pereira, F. C. N., & Warren, D. H. D. (1983). Parsing as deduction. In M. Marcus (Ed.),

Proceedings of the 21st Annual Conference of the Association for Computational Linguistics (pp. 137–144). Cambridge, MA: Association for Computational Linguistics.

Pickering, M. J., Traxler, M. J., & Crocker, M. W. (2000). Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language, 43,* 447–475.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: MIT Press.

Pinker, S. (1997). *How the mind works.* New York: Norton.

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

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39,* 347–370

Plaut, D. C. (1999). A connectionist approach to word reading and acquired dyslexia: Extension to sequential processing. *Cognitive Science, 23,* 543–568.

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes, 15,* 445–485.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Prasada, S., & Pinker, S. (1993). Similarity-based and rule-based generalizations in inflectional morphology. *Language and Cognitive Processes, 8,* 1–56.

Pullum, G., & Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19,* 9–50.

Rastle, K., & Coltheart, M. (1998). Whammies and double whammies: the effect of length on nonword reading. *Psychonomic Bulletin and Review, 5,* 277–282.

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science, 29,* 1007–1028.

Reali, F., Christiansen, M. H., & Monaghan, P. (2003). Phonological and distributional cues in syntax acquisition: Scaling up the connectionist approach to multiple-cue integration. In A. Markman & L. Barsalov (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 970–975). Mahwah, NJ: Lawrence Erlbaum.

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

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

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science, 29,* 819–865.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition, 72,* 68–109.

Roy, D. (2005). Grounding words in perception and action: Insights from computational models. *Trends in Cognitive Science, 9,* 389–396.

Samuel, A. C. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Expermental Psychology: General, 125,* 28–51.

Schütze, H. (1993). Word space. In S. J. Hanson, J. D. Cowan, & C. L. (Eds.), *Neural information processing systems 5*(pp. 895–902). San Mateo, CA: Morgan Kaufmann.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24,* 97–123.

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

Seidenberg, M. S., & Elman, J. L. (1999). Do infants learn grammar with algebra or statistics? *Science, 284,* 434–435.

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science, 23,* 569–588.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Seidenberg, M. S., & McClelland, J. L. (1990). More words but still no lexicon. Reply to Besner *et al.* (1990). *Psychological Review, 97,* 447–452.

Seidenberg, M. S., & Plaut, D. C. (1998). Evaluating word-reading models at the item level: Matching the grain of theory and data. *Psychological Science, 9,* 234–237.

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Smolensky, P. (1999). Grammar-based connectionist approaches to language. *Cognitive Science, 23*, 589–613.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind* (2 vols). Cambridge, MA: MIT Press.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word recognition down to the item level. *Psychological Science, 8*, 411–416.

St. John, M. F. (1992). The story Gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science, 16*, 271–206.

Steedman, M. (1999). Connectionist sentence processing in perspective. *Cognitive Science, 23*, 615–634.

Swier, R., & Stevenson, S. (2005). Exploiting a verb lexicon in automatic semantic role labelling. In R. J. Mooney (Ed.), *Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP-05) (pp. 883–890).

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.

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing. *Cognitive Science, 23*, 491–515.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 632–634.

Tomasello, M. (2003). *Constructing language*. Cambridge, MA: Harvard University Press.

Trueswell, J. C., & Tanenhaus, M. K. (1992). Toward a lexicalist framework for constraint-based ambiguity resolution. In J. C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 155–179). Hillsdale, NJ: Lawrence Erlbaum.

Wanner, E., & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. W. Bresnam, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 119–161). Cambridge, MA: MIT Press.

Weekes, B. S. (1997). Differential effects of number of letters on word and nonword latency. *Quarterly Journal of Experimental Psychology, 50A*, 439–456.

Williams, R. J., & Zipser, D. (1990). *Gradient-based learning algorithms for recurrent connectionist networks* (Tech. Rep. NU-CCS-90-9). College of Computer Science, Northeastern University.

Winograd, T. (1972). *Understanding natural language*. New York: Academic Press.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society, 104*, 444–466.

Zettlemoyer, L. S., & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence* (UAI-05).

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1131–1161.

The Cambridge Handbook of Computational Psychology

Edited by

RON SUN

Rensselaer Polytechnic Institute

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521674102

First published 2008

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data

The Cambridge handbook of computational psychology / [edited by] Ron Sun.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-85741-3 (hardback) – ISBN 978-0-521-67410-2 (pbk.)
1. Cognition. 2. Cognitive science. 3. Philosophy of mind. I. Sun, Ron, 1960–
II. Title: Handbook of computational psychology.
BF311.C36 2008
153.01'13–dc22 2007026278

ISBN 978-0-521-85741-3 hardback
ISBN 978-0-521-67410-2 paperback

CAMBRIDGE
UNIVERSITY PRESS