Chapter 15
Constraint Satisfaction Accounts of Lexical and Sentence Comprehension
Maryellen C. MacDonald and Mark S. Seidenberg

Consider what it takes to understand an ordinary sentence such as The man bought a tie with tiny white diamonds. Part of your understanding includes that the man is the agent of the action bought and that a tie is the thing being bought. To get this far, you also need to understand that man and tie are nouns and not verbs, although the verb usage of these words is possible in other contexts, such as man the boots or tie your shoes. You also need to understand that a tie refers to neckware, not to a game with equal scores, and that tiny white diamonds are an attribute of the tie, and not the currency used to purchase the tie (cf. The man bought a tie with his credit card). Despite these and many other possibilities, where interpretation could go wrong, the odds are in favor of your interpreting this sentence correctly. For example, man is more common as a noun than a verb, so a comprehender who unconsciously goes with the best odds will get to the right interpretation here. Similarly, words that follow determiners such as the and a are far more likely to be nouns than verbs, and tiny white diamonds are unlikely to be offered in trade for haberdashery, at least in most cultures. Comprehenders who follow the most likely alternatives will get to the correct interpretations of these aspects of the sentence. The idea that language comprehension is a process of following likely alternatives to derive an interpretation of ambiguous input form the basic claims of constraint satisfaction, or constraint based, theories of language comprehension. As in these examples, what is a likely alternative depends on properties of both individual bits of information (e.g., the frequency with which a word is used as a noun or verb and combinations of bits of information (e.g., the + man or a + tie)). Constraint-based theories emphasize how people learn, represent, and use such probabilistic information. This chapter will provide an overview of this approach, including its history, how it compares to alternative views, and a description of the kinds of computational mechanisms that are brought to underlie learning and using such constraints.

1. TRADITIONAL VIEWS OF LEXICAL AND SYNTACTIC AMBIGUITY
As the sentence about the man and his new tie illustrated, ambiguity is ubiquitous in language. This chapter will focus on two main types of ambiguity: lexical ambiguity, illustrated by the multiple meanings of words such as tie, and syntactic ambiguity,
illustrated by the alternative interpretations of *with tiny white diamonds* as something describing the tie (thus modifying a noun) or describing the method of buying (modifying a verb). (See the Pickering and van Greep, Kluender, and Tenenbaum and Trueswill chapters for discussions of other types of ambiguity.) The two kinds of ambiguity can interact; for example, adopting the noun *vs.* verb interpretation of *man* affects how one interprets the syntactic structure of a sentence containing this word. Despite the close relationship between these two types of ambiguity, for much of the history of modern psycholinguistics they have been studied independently. This division reflected differing views about lexical and syntactic representations (MacDonald, Pearlmutter, & Swendenberg, 1994). The meanings and other properties of words have often been thought to be stored in the *lexicon*, a person’s mental dictionary. On this view, interpreting words involves looking up, or *accessing*, information in the lexicon. This process was thought to be autonomous, proceeding in the same way regardless of the context in which a word occurred (Tenenbaum, Leiman, & Swendenberg, 1979; Swinney, 1979). It was also thought to make minimal demands on limited capacity working memory and attentional resources, allowing multiple meanings of words to be accessed in parallel. This led to a two-stage model of lexical ambiguity resolution. In the first stage, the lexical system accessed the common meaning or meanings of words; in the second stage, information derived from the linguistic and extra-linguistic contexts and the comprehender’s background knowledge were used to select the appropriate meaning and integrate it into the developing representation of the sentence (see Simpson, 1981, for review).

Syntactic structures, in contrast, were traditionally assumed to be *constructed* by a mental parser on the basis of grammatical rules. Deriving sentence structure was assumed to place demands on working memory and attentional resources that are limited in capacity (Frazier, 1987; Gibson, 1990; Just & Carpenter, 1992; MacDonald, Just, & Carpenter, 1992). These memory demands caused the parser to pursue only a single interpretation of syntactic structure at a time. This also led to a two-stage model. In the first stage, general parsing principles were used to assign a candidate syntactic structure online; in the second stage, other types of knowledge were utilized to flesh out this representation (e.g., interpret it semantically) and to revise the initial analysis if it were discovered to be incorrect.

Both lexical and syntactic accounts were motivated in part by appeals to the notion that language consists of distinct modules involving different types of information and processes; however, in the lexical case, this resulted in multiple alternatives being considered in a parallel process, whereas in the syntactic case, this resulted in a single interpretation being considered in a serial process (see MacDonald et al., 1994; Tenenbaum & Trueswell, 1995, for review). The modular view was consistent with distinctions between the lexicon and syntax in grammatical theories that were prominent at the time: the two-stage accounts were being developed (Newmeyer, 1980). The two-stage approach was also justified on the basis of assumptions about processing capacity limitations and the need to analyze the linguistic input very rapidly. The route to efficient interpretation was thought to be via a two-stage system in which the preliminary first-stage analysis prevented the input from being lost from working memory; the burden on working memory limitations was reduced because processing at this stage was limited to certain types of information.
e.g., syntactic structure. The initial interpretation could then be refined, corrected, and elaborated in the second-stage analysis. This attention to the time pressures of language comprehension and to the notion that processing may proceed through several distinct stages was reflected in the use of behavioral measures that were closely time-locked to the language input (e.g., tracking eye-movements, cross-modal priming). In various forms, two-stage approaches formed the dominant theoretical framework for word and sentence comprehension through the 1980s, and the focus on the time course of processing continues to this day.

The alternative view, which came to be called constraint-based language comprehension (or comprehension via probabilistic constraints), emerged in the 1990s. This approach challenged essentially every major tenet of the two-stage accounts. Whereas the two-stage theories held that comprehension consists of discrete stages at which different types of information and processes are used, constraint-based theories viewed comprehension as continuous and homogenous, with the same types of information and processes in use at all times. Whereas the two-stage theories assumed that processing limitations restrict the types of information that initially guide the comprehension process, constraint-based theories emphasized the richness of the linguistic signal, the capacity of language users to learn this information over time, and the comprehender’s capacity to bring this information rapidly to bear on the input during real-time comprehension processes.

2. SOURCES OF THE CONSTRAINT-BASED APPROACH

The constraint-based approach emerged from advances in several areas, including linguistic theory, corpus linguistics, psycholinguistics, and computational modeling.

2.1. Changing Views about Linguistic Structure

Whereas two-stage models reflect early approaches within generative grammar in which lexical and syntactic information were held to be separate, the constraint-based approach to comprehension is more closely related to work within linguistics in which (to varying degrees) lexical and syntactic representations are closely related (e.g., Breznahan, 1982; Chomsky, 1981; Joshi, 1985). The lexical representation of a word might include not only information about its spelling, pronunciation, and meaning(s) but also its grammatical functions and the types of syntactic structures in which it participates. It is a small step to then envision this information as part of a large interactive network (MacDonald et al., 1994). Under this scenario, the computation of both “lexical” and “syntactic” information in sentence comprehension is governed by a common set of lexical processing mechanisms.

1 This two-stage process is reminiscent of practices in machine translation, in which an automatic but limited first pass analysis (by machine) is then corrected and elaborated in a second-stage analysis (by a human translator).
2.2. Changing Views about Context

Language is comprehended essentially as it is perceived (Merslen-Wilson, 1975), and so a central question is what types of information can be brought to bear on decoding and interpreting the incoming signal. Studies of the role of language context in comprehension have also undergone a significant shift over the years. Research in the two-stage era focused on the use of real-world knowledge in guiding the comprehension process, and on the difficulties inherent in accessing relevant information online (Kintsch & Van Dijk, 1978). People know a vast amount about the world, as research on natural language processing in artificial intelligence suggested, it is a difficult problem to design a comprehension system that accesses relevant information from an enormous database of facts (Hayes-Roth & Jacobstein, 1994). Moreover, several studies emphasized the ineffectiveness of context, suggesting that comprehenders were limited in their application of real-world knowledge during comprehension (Forster, 1979), that context facilitated lexical processing only when words were highly predictable (Fischler & Bloom, 1979), and that this very strong degree of contextual constraint is rare in naturally occurring texts (Gough, Alford, & Holly-Wilcox, 1981). These results led to the conclusion that context-based prediction was not an important component of comprehension.

Complementary findings emerged from the study of lexical ambiguity resolution (e.g., Swinney, 1979; Tanenhaus et al., 1979). Many words are ambiguous between semantically distinct meanings (e.g., WATCH: a timepiece, to look; BANK: a monetary institution, the ground bordering a river). These early studies examined the processing of ambiguous words for which there are two male meanings that are used approximately equally often in the language ("equibased" ambiguities). The main finding was that subjects initially activated multiple meanings, even in contexts that were highly disambiguating. For example, the contexts in (1) and (2) clearly disambiguate the word ROSE. Yet subjects showed priming (facilitation compared to an unrelated control) for target words related to both of the main meanings (e.g., FLOWER, STOOD) presented immediately following each sentence (Tanenhaus et al., 1979). Results such as this were taken as evidence that comprehenders initially activated the common meanings of ambiguous words and within about 250 ms selected the correct meaning based on the context. Here too the processing of words seemed to be independent from processes involved in integrating a sequence of words into a meaningful, syntactically structured representation.

1. They all ROSE.
2. He bought a ROSE.

The research on predictability effects and lexical ambiguity resolution led many researchers to conclude that context effects are relatively weak, with the result that theories instead emphasized bottom-up aspects of processing—how words are identified. The ambiguity research played an important role in Fodor's (1983) development of his concept of modularity. The lexicon was seen as a paradigmatic example of an autonomous module in the comprehension system.
Subsequent research has led many of these conclusions to be revised. Whereas the
word predictability studies initially argued for a limited role of context, later work sugges-
ted that context effects could operate at levels other than predicting specific words.
Studies of semantic priming, for example, suggested that the processing of a word is
facilitated when preceded by a word with which it shares semantic features (e.g., McRae,
de Sa., & Seidenberg, 1997). Here the target words are not predictable, but facilitation
occurs nonetheless. Moreover, it is worth noting that most of the studies which suggested
that context effects are limited in scope examined reading rather than spoken language.
Written language does not exhibit many of the properties that make speech perception
such a difficult computational problem (e.g., variability with respect to rate, pitch, accent;
relatively lower signal-noise ratios; co-articulation and the absence of definitive mark-
ers for phoneme or word boundaries). The spoken code seems inherently more context
bound, insofar as the mere perception of sounds depends on the contexts in which they occur (e.g., Samuel & Pitt, 2003).

As with the context research, the lexical ambiguity research was similarly reexamined.
Whereas initial studies had argued for activation of multiple meanings of ambiguous words
independent of context, subsequent research yielded a more complex picture. Several stud-
ies showed that contextual information could result in only one meaning of an ambiguous
word being considered online (e.g., Simpson & Kreuger, 1991). However, other studies
showed that context could not override all aspects of lexical knowledge, in particular the
relative frequencies of the meanings: there was still an ambiguity effect (computation of
multiple meanings) when contexts favored the less-frequent meaning of an ambiguous
word (Duffy, Morris, & Rayner, 1988). Thus, the system is apparently neither strictly mod-
ular nor completely context-bound. Kawano and Kuroda (1993) developed a computational
model that provided insight about these results. His system was not inherently modular, insofar as
nothing architectural prohibited contextual information from affecting meaning activation.
However, in practice lexical information was activated more rapidly, limiting the effects of
context. This is because there is a much closer relationship between the spelling or sound
of a word such as ROSE and its meanings than there usually is between either of the mean-
ings and the contexts in which they occur.

Finally, researchers began to question a key assumption underlying much of the
research on lexical ambiguity: that words have discrete meanings that can be accessed
like entries in a mental dictionary. The meaning of a word routinely shifts as a function of
the context in which it occurs. Consider a word such as piano. It has a seemingly
simple, unambiguous common meaning: large keyboard instrument with steel wires
struck by felt-covered hammers (we are ignoring here the secondary musicoacoustical
sense meaning "soft in volume"). Yet different shades of meaning are involved in push-
ing a piano (where weight is relevant but musical properties are not) vs. playing a piano
(where the opposite is true; Merrill, Sperber, & McCauley, 1981). How to properly
categorize meanings is a difficult issue that has been addressed from many theoretical
and disciplinary perspectives (Margolis & Laurence, 1999). Here it is sufficient to note
that it may be an essential property of word meaning that it is conceived in a context-
dependent manner every time a word is comprehended. This type of computation
seems inherently at odd with a modular lexicon that automatically and indiscriminately
activates stored meanings and passes them along to other comprehension systems. The creation of novel meanings from proper nouns (Clark & Clark, 1977) and the interpretation of novel noun compounds (Gagne & Slobin, 1977) raise similar issues.

We do not have a general theory of lexical ambiguity resolution in hand; to have one would be to solve a good part of the problem of language comprehension. However, this research made it clear that a broad range of factors involving properties of both words and contexts affect lexical ambiguity resolution, and that the interactions among these many factors determine the outcomes that are observed.

2.3. Changing Views about Language Statistics

Languages exhibit statistical structure—variations in the distributions of elements such as sounds, words, and phrases. Despite the existence of this structure, many years of statistical analyses of language attract little interest within mainstream linguistics and psycholinguistics, principally because Chomsky (1957) compellingly argued that language exhibits important properties that are not captured by mere statistics (as “Colorless green ideas sleep furiously” illustrated). According to the probabilistic constraints approach, however, comprehension essentially is the process of exploiting statistical regularities of many kinds. Learning and using language seem like difficult problems (ones that necessarily require innate grammatical knowledge, or learning or parsing mechanisms) only because this statistical information was systematically excluded from theorizing.

The ground-breaking studies that expanded notions about the range of information that might be used in sentence comprehension were Bever (1970) and Ford, Brennan, and Kaplan (1982). In a classic article, Bever (1970) made a number of important observations concerning syntactic complexity and ambiguity and the factors that make sentence comprehension difficult. Bever suggested that comprehenders are guided by perceptual strategies that assign interpretations based on frequency and plausibility. He described a specific strategy whereby comprehenders interpret noun-verb-noun sequences as agent-action-object. Violating this expectation (as in Bever’s example “The horse raced past the barn fell”) creates a misanalysis, which came to be known as a “garden path” effect (Frazier, 1978). Ford, Brennan, and Kaplan (1982) provided an early investigation of the effects of lexical knowledge on sentence comprehension. They proposed that comprehenders initially adopt an analysis of a syntactic ambiguity that incorporates the most frequent subcategorization of the sentence’s verb (see also Fedor, 1978). Verb subcategorization refers to the noun phrase arguments a verb may take; for example move may or may not have a direct object noun phrase. Ford et al. provided evidence consistent with the idea that the several subcategorization options were ordered by frequency, and that comprehenders consider sentence interpretations in the corresponding order.

Although their importance was widely recognized, the Bever and Ford et al. articles did not immediately generate a program of research. One problem that inhibited further progress was that the research tools that were available did not make it easy to calculate
robust language statistics from large samples of text or discourse. This problem was largely obviated in the 1990s, when resources such as the Wall St. Journal corpus (Marcus, Santorini, & Marcinkiewicz, 1993) became publicly available and could be analyzed using desktop computers. This methodological advance made it possible to conduct behaviora1 studies examining the use of various types of statistical information in comprehension (discussed below). A second problem was the absence of a theory that could explain which language statistics are relevant, and how they could be learned, represented in memory, and efficiently used in processing. Is the absence of such a theory, it was not obvious how the Bever and Ford et al. findings could be extended. This problem also began to be addressed in the 1990s, with advances in the theory of statistical learning within the connectionist framework, to which we now turn.

2.6 Development of the Connectionist Paradigm

The term “connectionist” refers to a broad, varied set of ideas, loosely connected (so to speak) by an emphasis on the notion that complexity, at different grain sizes or scales ranging from neurons to overt behavior emerges from the aggregate behavior of large networks of simple processing units. Our focus is on the parallel distributed processing (PDP) variety developed by Rumelhart, McClelland, and others in the 1980s (McClelland, Rumelhart, & Hinton, 1986). This approach includes a variety of concepts that are potentially relevant to language. In brief, PDP networks consist of large numbers of simple processing units that take on activation values. The connections between units carry weights that determine how activation is passed between units. The network is configured to perform a task (such as recognizing a word or object, or predicting the next word in a sentence). Learning involves gradually adjusting the weights on connections.

The problem is to find a set of weights that yields performance that corresponds to human performance on the task (e.g., with respect to accuracy, generalization, developmental trajectory). Several algorithms can be employed for this purpose; they vary in how closely they mimic properties of learning at neural or behavioral levels (see Harm & Seidenberg, 2004, for discussion). Network performance is determined by several main factors: (1) the architecture of the system (e.g., the configuration of units and connections); (2) the characteristics of the input and output representations; (3) characteristics of the patterns used in training the model; and (4) characteristics of the learning algorithm. In other words, the model's performance depends on its initial state, what it experiences, and how it learns from those experiences.

This theoretical framework has been discussed extensively elsewhere; here we focus on three properties that inform the probabilistic constraints approach to comprehension.

First, the networks incorporate a theory of statistical learning. The main idea is that one way that people learn (there may be others) is by gathering information about the frequencies and distributions of environmental events. This type of learning is thought to be general rather than language specific. Many nonhuman species are also capable of rudimentary forms of statistical learning (Estes, 1955); humans may be distinct with respect to the power of their statistical learning capacities. Language, for example, requires
tracking correlations and covariation across multiple types of linguistic information within and across modalities (e.g., a speech signal and the context in which it is uttered).

The applicability of these ideas to language was initially explored in the context of learning inflectional morphology (Kempehart & McClelland’s, 1986, past tense model) and learning to read (Seidenberg & McClelland, 1989; Plaut McClelland, Seidenberg, & Patterson, 1996; Harm & Seidenberg, 2004). The reading models in particular developed the idea that lexical knowledge consists of statistical relations between orthographic, phonological, and semantic codes. Learning then involves acquiring this statistical knowledge over time. Subsequent research on statistical learning in infants and adults has provided strong evidence consistent with this view. A wealth of studies now attest to humans’ robust abilities to learn statistical patterns that inhere in diverse types of stimuli (Saffran & Sahin, in press). The domain-generality of statistical learning is suggested by studies showing that infants are equally good at learning the statistical structure in a series of spoken syllables and a series of pure tones (Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999), and by similarities across auditory (Saffran et al., 1996) and visual (Kirkham, Stemmer, & Johnson, 2002) modalities. This learning mechanism provides a way to derive regularities from relatively noisy data, a property that is likely to be highly relevant to the child’s experience in learning language. Although some researchers argue that specifically grammatical relationships are not acquired by statistical learning (e.g., Marcus et al., 1999; Peña, Bonatti, Nespor, & Mehler, 2002), these claims have been challenged (Pernacca, Tyler, Galland, & Peereman, 2004; Seidenberg, MacDonald, & Saffran, 2002).

Second, the models provide a basis for understanding why particular types of statistics are relevant and not others. Above we described the main factors that determine a model’s behavior (and, by hypothesis, a person’s). Note that this description did not include a specification of which types of statistics a model should compute. It is not necessary to stipulate this in advance because this aspect of a model’s behavior falls out of the other factors. In practice, what a model learns is heavily determined by the nature of the representations that are employed. These representations (e.g., of phonology or semantics) are intended as (simplified) claims about what people know and bring to a task such as language learning. This knowledge may be innate or may itself be learned by processes to be explored in other models. The goal is to endow a model with exactly the knowledge and capacities that people (infants, children, adults) bring to learning a task, although this ideal is only approximated in any implemented model. Given the properties of these representations, other aspects of the model architecture (e.g., number of units or layers; patterns of connectivity between layers), and a connectionist learning algorithm, the model will pick up on particular statistical regularities implicit in the examples on which the model is trained. Thus, motivating the various elements of a model and how it is trained is very important, but the model itself determines which statistics are computed.

This discussion is relevant to a concern that is often voiced about connectionist models, that they are too powerful – capable of learning regularities that humans cannot learn. In fact, what such models learn is highly constrained. Constraints on what is
learned arise not so much from the learning algorithm itself as from other aspects of the network, particularly properties of the representations that are used. For example, models that represent articulatory or acoustic primitives in a realistic way are constrained by facts about what people can say or hear.

Third, the framework provides a powerful processing mechanism, the exploitation of multiple simultaneous probabilistic constraints. Information in a network is encoded by the weights. The weights determine ("constraint") the output that is computed in performing a task. Processing involves computing the output that satisfies these constraints. This output changes depending on what is presented as input (e.g., the current word being processed in a network that comprehends sentences word by word).

This type of processing, known as constraint satisfaction, has several interesting properties. One is that the network's output is determined by all of the weights. Such models illustrate how a large number of constraints can be utilized simultaneously without imposing excessive demands on memory or attention. Constraint satisfaction is passive—activation spreads through a network modulated by the weights on connections—rather than a resource-limited active search process. Another important property is that the constraints combine in a nonlinear manner. Bits of information that are not very informative in isolation become highly informative when taken with other bits of information. Much of the power and efficiency of the language comprehension system arises from this property. Languages exhibit many partial regularities. Different types of information are correlated, but weakly. The comprehender cannot wholly rely on any one type of information because combinations of these partial cues are highly reliable. This concept may seem paradoxical at first. If individual cues are unreliable, wouldn't combinations of these cues be even more unreliable? No, not if cue combination is nonlinear. The informativeness of each cue varies as a function of other cues. This point is easy to grasp by illustration. Someone is thinking of an object—guess what it is. The cues are it is a fruit, it is yellow, and its name begins with B. In isolation, each cue only weakly constrains the answer. The combination of cues, however, makes it very likely that the object is banana.

The same process can occur on a sentence or discourse level. In the context of a discussion of shopping and the syntactic environment of the determiner a, the word tie probably refers to neckwear. This contingency holds despite the fact that all the simple probabilities are quite low—by itself, a shopping context doesn't demand that neckwear be discussed, the occurrence of a does not predict the word tie, and tie in isolation affords several more frequent interpretations than the neckwear one.

The bases of constraint satisfaction systems have been explored extensively in the computational literature. Connectionist models provide one way of implementing this process, but there are symbolic systems that perform similarly (MacWhinney, 1977). In the psycholinguistic literature, the basic idea was introduced in Bates and MacWhinney's (1985) Competition Model. Bates and MacWhinney argued that language is comprehended by following "cues" that compete with one another and are weighted as a function of their effectiveness in past comprehension events. The Competition Model incorporated
the important ideas that many linguistic cues are learned and language-specific; that cues could conflict and be differentially weighed; and the importance of integrating syntactic and nonsyntactic information during comprehension. The probabilistic constraint approach can be seen as coupling many ideas embodied by the Competition Model with proposals about the statistical basis of cues ("constraints"); and how multiple constraints are learned, represented, and exploited in processing. The Competition Model has been very important in research on how children acquire knowledge of language-specific cues, how languages differ with respect to the relative prominence of different cues (e.g., word order vs. inflectional morphology), and how cue competition affects the final interpretation of a sentence. The model had less to say about the integration of many simultaneous probabilistic cues, or about online processes in comprehension (see Elman, workspace, and McRae, 2004, for discussion). Also, in the connectionist models we have described, different alternative interpretations do not directly compete. The same weights are used in processing all input patterns. The performance of the model (or person) depends on the aggregate effects of exposure to many examples. There is nothing like parallel activation of multiple alternatives, just the computation of the best-fitting output. "Competition" is realized only implicitly, because alternatives have affected the weights, not by explicitly computing and comparing alternatives.

In summary, the probabilistic constraints approach emphasizes the role of statistical information concerning the occurrence and co-occurrence of different types of linguistic and nonlinguistic information in language comprehension. Learning a language involves acquiring this information from the large sample of utterances to which every learner is exposed. The theory assumes that humans are born with (or soon develop) capacities to perceive particular kinds of information (e.g., in listening), to engage in statistical learning, and to encode what is learned in networks of neurons. Familiar types of linguistic representations such as phonemes, syllables, morphemes, words, and constituents are not represented directly in memory; rather these terms are approximate descriptions (or higher level statistical generalizations that emerge with experience (e.g., Seidenberg & Gordon, 2000). On this view, the learner builds (or "bootstraps") a language out of statistical relations among different types of information, and skilled language comprehension involves using these statistical generalizations in processing utterances. These ideas have been extensively explored in the context of syntactic ambiguity resolution, to which we now turn.

3. PROBABILISTIC CONSTRAINTS AND SYNTACTIC AMBIGUITY RESOLUTION

Syntactic ambiguities arise when a sequence of words is compatible with more than one sentence structure. Often the syntactic ambiguity coincides with a lexical ambiguity of some sort. For example, in (3), there is an ambiguity between interpreting Carol as the noun phrase (NP) direct object of the verb saw or the beginning of a sentential complement (often termed the NPS ambiguity). This ambiguity is linked to lexical ambiguity in the verb, which can optionally take either a direct object NP or a sentential complement. The example also illustrates another common feature of syntactic ambiguities, at least in
English, that they may be triggered by the omission of an optional word or phrase. Thus, the sentential complement sentence in (1c) could be introduced with that (Shanta saw that Carol ...), which would remove the temporary ambiguity.

3a. Temporary ambiguity: Shanta saw Carol ...
3b. NP direct object interpretation: Shanta saw [Carol], but Carol didn’t see her.
3c. Sentential complement interpretation: Shanta saw [Carol would be late].

A dominant concern in syntactic ambiguity resolution has been the timecourse over which information is brought to bear on the ambiguity. The modular two-stage account is exemplified by Frazier and colleagues’ Garden Path Model (e.g., Frazier, 1987; Frazier & Clifton, 1996), in which the first-stage parser (the syntactic interpretation component of the comprehensive system) develops a syntactic structure for the input, guided by only the lexical categories of the input words (noun, verb, etc.), the syntactic rules of the language, and by structure-based heuristics (Minimal Attachment and Late Closure) that direct structure building when more than one alternative structure is afforded by the input. At some later point, a second stage integrates semantic and contextual information into the representation, and if this information conflicts with the initial interpretation built by the parser, the conflict may trigger a revision and reanalysis of the input.

The constraint-based view argues that the preference for one interpretation over another during comprehension of an ambiguous sentence stems not from global heuristics such as Minimal Attachment but from the rapid combination of many probabilistic constraints. A key observation concerning such constraints is that different types of information tend to be correlated; for example, a verb’s meaning is strongly related to the kinds of noun phrases it tends to appear with in sentences (Hare, McRae, & Elman, 2003; Levin, 1993; Roland & Jurafsky, 2002). As a result, even weak cues can combine with other correlated cues and have a strong effect on interpretation preferences. Thus the approach links syntactic level information, such as knowledge about transitive sentence structures (those with a direct object in the verb phrase), to lexically specific information, such as the frequency with which a particular verb (bought, say) occurs with a direct object and the frequency with which a noun (e.g., tie) occurs as a direct object, and the conjoint frequency with which bought and tie co-occur in a verb phrase with买的 object configuration. The correlation of cues has an important role in understanding how abstract pragmatic constraints, often thought to be too complex to be brought to bear in online ambiguity resolution, could have a rapid effect on the process. For example, new entities introduced into a discourse are more likely to receive additional modification than are previously mentioned (or “given”) noun phrases, thus affecting the probability that a syntactically ambiguous prepositional phrase will modify this noun phrase. The given/new distinction is strongly correlated with the type of determiner used to introduce the noun phrase; new entities often occur with a, and given ones with the. Thus a tie with is more likely to have the with phrase modify the tie than is the sequence the tie with ... (Spivey-Knowlton & Sedivy, 1995). This pattern reflects how discourses are structured and might require extensive computation in some cases, but the comprehender has a ready proxy in the simple co-occurrence of some determiners and the interpretation of with.
The contract between two-stage and constraint-based accounts has often focused on the extent to which the separate processing stages posited by two-stage models are isolable. In the case of the Garden Path model, in which a purely syntactic first stage is followed by use of all other types of information in a second stage, the issue in the extent to which putatively second-stage nonsyntactic information could be shown to affect the operations of the first-stage parser. A significant body of work in the 1980s and 1990s used eye fixations during reading to address this issue, and a number of researchers suggested that the earliest eye fixations on a small region of text reflected operations of the first-stage parser, while later fixations were driven by second-stage semantic integration processes (e.g., Rayner, Carlson, & Frazier, 1983). This view was motivated in part by studies in which manipulations of semantic information in syntactically ambiguous sentences were found to affect late eye fixations, but not early ones (e.g., Ferreira & Clifton, 1986; Rayner et al., 1983). Subsequent studies suggested that the delayed effects of nonsyntactic information in these reading patterns were attributable to weak or infective contexts or other biases in the ambiguous stimuli (Altman & Steedman, 1988; Crain & Steedman, 1985). More robust manipulations of context have shown clear evidence of the use of nonsyntactic information in first pass reading measures (e.g., Garney, Pearlman, Myers, & Lotzcky, 1997; Trueswell, Tanenhaus, & Garney, 1994) and even in the first fixation durations on words (Speer & Clifton, 1998), a measure that has often been taken as the earliest processing evidence obtainable with eyetracking (Rayner et al., 1983). Moreover, as the nature of contextual effects received additional investigation, the number of potentially relevant constraints, and the interactions between them, grew more complex. This trend can be illustrated by considering one particular ambiguity in detail.

3.1. An Example: Main Verb vs. Reduced Relative Ambiguities

The structures considered here are probably the most thoroughly studied in psycholinguistics. The focus on these structures arose from Bever’s (1970) observation that whereas the sentence *The horse raced past the barn fell* is taken to be gibberish by most speakers of English, it is readily comprehended when two optional words (a relative pronoun and a form of be, such as *that was*) are inserted marking the start of a relative clause, as in *The horse that was raced past the barn fell*. Another example, somewhat easier to comprehend, is given in (4). This is called the Main Verb/Reduced Relative (MVRR) ambiguity because it is initially unclear whether the first verb, *raced* in Bever’s example and *arrested* in (4), is the main verb of the sentence (as in 4b) or is introducing a reduced relative clause (4c). The clause is said to be “reduced” because of the omission of the optional relative pronoun and a form of be.

4a. Temporary Main Verb/Reduced Relative Ambiguity: The three men arrested...

4b. Main Verb Interpretation: The three men arrested the burglary suspects in a parking garage.

4c. Reduced Relative Interpretation: The three men arrested by the local police were wanted in connection with the jewel robbery.
Early studies of interpretation of thin ambiguity manipulated the degree of contextual information consistent with the reduced relative (RR) interpretation and found strong misinterpretation of "garden-path" effects in reading patterns at all levels of contextual support, indicated by long reading times in the sentence region that disambiguated the ambiguity (Rayner et al., 1983; Ferreira & Clifton, 1986). These reading patterns were taken to indicate that comprehenders initially adopted the main verb (MV) interpretation (the one favored by the parsing heuristic: Minimal Attachment) independent of context and were surprised when the disambiguation favored the reduced relative interpretation. Subsequent studies explored the nature of contexts in depth and suggested that interpretation is guided by a number of probabilistic constraints, with the difficulty of a given interpretation of the ambiguity typically a function of several constraints acting together. Some of the major categories of constraints are listed below.

A. Animacy of the pre-verbal NP (e.g., men), as this affects the likelihood that this noun will be the agent vs. patient of an upcoming verb, in that animate nouns are more typical agents. This constraint is important because the noun is the agent of the next verb in the MV interpretation, and it is the patient of the verb in the RR interpretation (Trueswell et al., 1994), but Ferreira & Clifton, 1986 failed to find animacy effects.

B. The relative frequency of usage of the ambiguous verb (e.g., crossed) in active vs. passive structures, as the MV interpretation is an active structure while the RR is a passive. Active/passive voice frequency is related to several intercorrelated properties of the verb, including the verb's frequency of occurrence in the past tense (required for the active MV interpretation) vs. past participle (required for passives and the RR interpretation), and its relative frequency of use in transitive (with a direct object) vs. intransitive (no direct object) constructions. The RR interpretation is always transitive, but the MV may be intransitive (Hare, Tanenhaus, & McRae, 2006; MacDonald, 1994; MacDonald et al., 1994; Trueswell, 1996).

C. Plausibility of the pre-verbal NP as an agent vs. patient of the ambiguous verb, such as the plausibility that men would be the agent vs. patient of crossed, what McRae, Ferretti, & Amyotte (1997) called thematic fit. It is an example of a combinatorial constraint, in that it integrates properties of at least two words and the information in constraints A-B above (McRae, Spivey-Knowlton, & Tanenhaus, 1998; Pearlmuter & MacDonald, 1992; Tagoni, Spivey-Knowlton, McRae, & Tanenhaus, 1994). The power of such combinatorial constraints can be seen in several reanalyses of failures rapid effects of some simple constraint. For example, studies that found to find only minimal effects of noun animacy or other broader discourse plausibility factors (Ferreira & Clifton, 1986; Rayner et al., 1983) tended to use stimuli with verbs that strongly promoted an active, intransitive interpretation. In this situation, verb biases were working strongly in favor of the MV interpretation, and combinatorial constraints (over properties of both verbs and nouns) had little effect (MacDonald et al., 1994; Tanenhaus & Trueswell, 1995; Trueswell et al., 1994).

D. The basic frequency of the MV vs. RR structure. Within two-stage models, the initial preference for MV structures stems from parsing heuristics such as Minimal Attachment, but within the constraint-based tradition, this effect emerges from the
fact that MV structures are more common than RR structures in the language (Berger, 1970; McRae et al., 1998).

E. The nature of the words after the onset of the ambiguity. In most empirical studies, the first few words after the ambiguous verb constitute a prepositional phrase, such as by the local police in sentence (4c) above. Depending on their lexical properties and that of the preceding material, the post-ambiguity words may serve to promote one or the other interpretation of the ambiguity. The constraints here can be simple, such as the basic probability that by refers to an agent of action (promoting the passive and thus an RR interpretation) vs. a location (less constraining for the two alternative interpretations), or the constraints may be combinatorial, such as properties of by given a particular preceding verb or NP, as in by the local police (MacDonald, 1994; McRae et al., 1998). Following the prepositional phrase, the relative clause typically ends in most stimulus materials, and the true main verb of the sentence is encountered, as in we were wanted in France. Researchers often assume that encountering the main verb completely disambiguates the string in favor of the RR interpretation, but the degree of disambiguation actually varies greatly with particular stimulus items. The major factor here is whether the main verb is itself ambiguous between a past tense and a past participle interpretation. A tense ambiguity at the main verb permits a second temporary MV/RR ambiguity in the stimulus sentence, as in The witness examined by the lawyer turned out to be unreliable (from Ferreira & Clifton, 1986). Here turned initially permits a RR modifying lawyer (as in the lawyer turned in by the detective), so that a definitive disambiguation is delayed. Stims with this second ambiguity are rare in most studies (including in Ferreira & Clifton), but they may be a source of additional noise in the reading data in some experiments. This additional ambiguity also serves to reinforce the point about the large number of constraints that can influence ambiguity resolution here.

F. The thematic role of the pre-verbal noun. Relative clauses are more natural when the head noun is a theme of the action (the flowers that were sent to the performer) than when the recipient of the action is the head noun (the performer who was sent the flowers...) (Keenan & Comrie, 1977). The typical (or oddness) of modifying a Goal role decreases the likelihood of a reduce relative interpretation, and some studies that have found poor use of nonsyntactic constraints have failed to contain stimuli in which the goal role is relativized (e.g., Rayner et al., 1983), which strongly promotes the MV interpretation. This bias also interacts with the effect of post-verb words described in point E, in that when the goal NP is modified as in the performer sent the flowers, the words after the ambiguous verb (the flowers) can strongly promote the MV interpretation (Tabor, Galantucci, & Richardson, 2004).

G. Constraints from the broader discourse that could promote either interpretations. These constraints include whether the discourse makes it plausible to modify the first noun, which promotes an RR interpretation (Altman & Steedman, 1988; Crain & Steedman, 1985; N., Crain, & Shankweiler, 1996; Sedivy, 2002), whether the tense of the verbs in prior discourse promotes interpretation of the ambiguous verb as a past tense or past participle (Trueswell & Tanenhaus, 1991), and factors affecting the likelihood of using a reduced vs. unreduced relative clause form in various discourse situations (McKoon & Ratcliff, 2003). The influence of these
discourse-level constraints may be modulated by more robust lexical-level constraints. For example, Filik, Paterson, and Liversedge (2005) found the extent to which attention-focusing words such as only influenced ambiguity resolution (Ni et al., 1996; Sedivy, 2002) varied with the range of alternative interpretations permitted by the ambiguous verb.

This and similar tests of potential constraints and their interactions (Townsend & Bever, 2001; Tennerhau, Spivey-Knowlton, & Hanna, 2000) suggest why comprehension of the MV/RR ambiguity can sometimes succeed easily and other times fail miserably. Successful comprehension occurs when a variety of constraints strongly promote the ultimately correct reduced relative interpretation at an early point is the ambiguous string, and garden-pathing occurs when the evidence strongly favors the incorrect interpretation, as in The horse raced past the barn fell.

3.2. Computational Modeling of Constraint-Based Theory

The probabilistic constraints approach draws heavily on computational concepts such as constraint satisfaction. One way of exploring such concepts is by implementing simulation models. Modeling is a tool that can be used for different purposes; here we describe three ways computational models have been used in the development of the constraint-based approach.

1. As a tool for developing and illustrating novel theoretical concepts and analyses of the comprehension process. These models tend to be narrow in scope and tied to phenomena rather than the results of particular behavioral studies. Perhaps the most influential example is the work of Elman (1990), who developed the concept of a simple recurrent network in which the task is to predict the next word in a sentence given the current word and information about the prior context (in Elman’s networks, the state of the hidden unit layer). Elman’s models exhibited several interesting behaviors: they learned to predict words that were grammatical continuations of sentences; they formed representations of the grammatical categories of words; they encoded long-distance dependencies, not merely transition probabilities between adjacent words. The models introduced the important idea that sentence comprehension could be construed as following a trajectory in the state-space defined by a recurrent network. The models also helped to revive the idea that prediction might be an important component of sentence interpretation; recall that early results suggesting that words are not generally predictable from context (Gough et al., 1981) led to the view that contextual effects were weak and unhelpful. However, Elman’s networks, and other connectionist approaches that emphasized distributed representations, suggested that comprehension processes might incorporate partial predictions where expectations are generated for certain semantic or syntactic properties of the upcoming input, even if exact words themselves are not predicted. There is now increasing evidence that human comprehension processes incorporate these predictive elements (e.g., Altmann, van Nice, Gurnham, & Henstra, 1998; McRae, Hare, Elman, & Ferretti, 2005).
The St. John and McClelland (1990) model combined an Elman network with a distributed hidden layer of sentence meaning representations (the “sentence ginal”). The model took simple sentences as input and was trained to answer queries about the thematic roles of noun phrases. The fact that the model was trained to represent meaning rather than predict upcoming words made it an interesting departure from other models of sentence processing, but it was very limited in scope, having a small number of words and thematic roles. The model also could not interpret multi-clause sentences, a serious limitation given the centrality of these constructions in syntactic and psycholinguistic research. Robale (2002) adapted and expanded the St. John and McClelland architecture in a much larger model. His model replicated several key results in human sentence interpretation, but also behaved in some ways that differed from humans. Further research would be needed to determine whether these limitations could be overcome within this architecture.

A final example is Allen and Seidenberg’s (1999) model in which both comprehension and production were simulated within a single network. They used the model to illustrate a theory of how people make grammaticality judgments, and how, paradoxically, this ability could be maintained in aphasia (Linberger, Schwartz, & Saffran, 1983). The model also illustrated why a sentence such as “Colorless green ideas sleep furiously” is judged grammatical even though the transition probabilities between words are low.

None of these models are “complete” in any sense, and the limitations on their scope allow the possibility that the results are not general. Nonetheless, these models are important as vehicles for introducing novel mechanisms, approaches, and analyses.

2. As a procedure for discovering the statistical regularities implicit in a large corpus of utterances, as discussed above. This application of modeling has been used mainly in studies of language acquisition; studies by Minn (2003), Redington, Chater, and Finch (1998) and others show how representations of grammatical categories can be derived from distributional information. Cassidy, Kelly, and Sharoni (1999) used a simple feedforward network as a procedure for discovering phonological correlates of proper names. Haskell, MacDonald, and Seidenberg (2003) used a network to discover phonological properties associated with adjectival modifiers in English. Mitrović, MacDonald, and Seidenberg (2005) developed a model that learned much of the complex inflectional system for nouns in Serbian, and showed that gender was cued by correlations between phonology and semantics.

3. As a way of accounting for behavioral data. This usage is akin to models of word reading that simulate the results of behavioral experiments (e.g., Harm & Seidenberg, 2004). Some simple recurrent networks (SRNs) have been used for this purpose. These SRNs are typically subject to additional analyses that allow their behavior to be linked to measures of human performance (Christiansen & Chater, 1999; MacDonald & Christiansen, 2002; Tabor, Juliano, & Tanenhaus, 1997; Tabor & Tanenhaus, 1999). For example, MacDonald and Christiansen computed an error
measure, called Grammatical Prediction Error, that provided a good fit with reading times in studies of relative clause comprehension. Tabor, Juliano, & Tanenhaus, (1997) and Tabor and Tanenhaus (1999) related model performance to reading times using a different approach. They coupled an SRN and a dynamical processor that was designed to relate hidden unit representations to a sentence interpretation. The dynamical processor’s ability to settle on an interpretation varies with past experience, and processor time can be related to reading times in behavioral studies.

Another class of models directly addresses the process of constraint-weighting during comprehension and the linkage between these processes and behavioral data (Elman et al., 2004; McRae et al., 1998; Spivey & Tanenhaus, 1998). These models are not simulations of sentence comprehension per se but rather test claims for activation of alternative interpretations at different points in a sentence. Moreover, the models have no learning component but are instead hand-tuned to the properties of particular stimulus sentences from one or more existing behavioral study, potentially yielding a transparent relationship between the activation levels in the model and the pattern of behavioral data. For example, McRae et al. modeled the interaction of six constraints in the MV/RR ambiguity. They developed stimulus materials for an empirical study and used a combination of corpus analyses and questionnaire studies to assess the degree of constraint promoting MV vs. RR interpretations for each of their stimulus items. The constraints were implemented in a simple localist model in which each constraint and each interpretation of the ambiguity are represented by single nodes in a network. The model simulates the timecourse of constraint interaction because constraints are combined at each word position, and the alternative interpretations receive activation as a function of the combined constraint strength at that point. The alternative interpretations compete with one another, so that activation of one alternative drives down activation of the other. McRae et al. used the model to test two alternative accounts of ambiguity resolution for this structure by varying the time at which different constraints were available to the model. In one model, all syntactic and nonsyntactic constraints were available as soon as the relevant words were encountered in the sentence; this model corresponded to the claims that ambiguity resolution is accomplished through the rapid integration of multiple probabilistic constraints. In the second model, a syntactic constraint favoring the MV interpretation was allowed to have an early effect, and non-syntactic constraints (verb tense frequencies and the plausibility of a noun being an agent or patient of a verb) were delayed for several words. This model was designed to simulate predictions of a two-stage model, in which the MV interpretation is adopted in the first stage and use of non-syntactic information is delayed until the second stage. McRae et al. found that the more interactive model was a better fit to the data than the one in which non-syntactic constraints were delayed. They argued that the addition of modeling provides a much stronger test of alternative accounts than empirical work alone, in that the modeling effort forces commitments to particular claims about constraint interaction and its timecourse.

This brief summary serves to show that a wealth of ideas about sentence comprehension and related aspects of language have been explored using implemented connectionist models.
4 STATE OF THE SCIENCE: CONTROVERSIES, UNRESOLVED ISSUES, AND FUTURE DIRECTIONS

The approach we have described is ambitious and yet still in the early stages of development, and so there are many gaps that researchers are attempting to address. We therefore conclude this chapter by considering a series of questions.

4.1 Statistics All the Way Down?

A number of researchers have taken issue with constraint-based approaches to language comprehension processes (e.g., Frazier, 1995, 1998; Townsend & Bever, 2001). We have discussed some of these concerns elsewhere in this chapter, and our focus in this section will be on empirical studies that are designed to provide evidence that important aspects of language comprehension have a nonstatistical basis, contradicting a basic tenet of the approach. For example, McKoon and Ratcliff (2003, 2005) have challenged the constraint-based account of ambiguity resolution in the MV/RR construction, and more broadly, suggested that difficulty with RR sentences such as *The horse raced past the barn fell*, lie not in their temporary ambiguity but in the incompatibility between the construction and the verb’s meaning (specifically, that *race* is a verb with internally caused changes of state). They suggest that the RR sentences and their “unreduced” counterparts, such as *The horse that was raced past the barn fell*, have subtly different meanings and uses, such that internally caused change of state verbs can appear in unreduced but not reduced relative clauses. This approach, in which certain reduced relatives are nonsensical rather than merely ambiguous, to some degree indicts all ambiguity resolution approaches to this construction. McKoon and Ratcliff’s claims have been forcefully countered by McRae, Hare, and Tanenhaus (2005) and Hare et al. (in press). They trace the difficulty in McKoon and Ratcliff’s examples to the frequency of passive uses of the ambiguous verbs (see constraint B in the list above), disentangle this property from meaning components, and provide additional evidence for a constraint-based account of this ambiguity.

Perhaps the most direct empirical challenge to the constraint-based accounts comes from work by Pickering, Traxler, Van Gompel, and colleagues (e.g., Traxler, Pickering, & Clifton, 1998; Van Compel, Pickering, & Traxler, 2001; Van Gompel, Pickering, Pearson, & Livermore, 2005), who have argued that constraint-based accounts make incorrect predictions about reading times for certain ambiguities. Specifically, they observe that constraint-based accounts predict that comprehension times should be longer for ambiguous sentences compared to unambiguous ones, owing to the fact that ambiguous sentences engender competition between alternative interpretations. In a series of studies using several different ambiguous constructions, these authors have found that globally ambiguous sentences are read more quickly, not more slowly, than unambiguous sentences. They suggest that these data argue against a constraint-based account and instead support a two-stage model in which multiple sources of information may affect a sentence’s initial interpretation. However, Green and Mitchell (2006) found that the McRae et al.’s (1998) computational model generally did not enter into an extended (period of competition for globally ambiguous sentences, and thus there is no prediction
CHAPTER 15. CONSTRAINT SATISFACTION ACCOUNTS

for longer reading times for these items relative to disambiguated ones in these models. Green and Mitchell’s simulations uncover behavior in the model that runs contrary to the assumptions that Tricker et al. (1998) and Van Gompel et al. (2001, 2005) made about model performance, and their simulation results emphasize the pitfalls of relying on intuition for how an implemented model will behave. More generally, these behavioral results and simulations will serve to push alternative accounts to be more precise, especially about predictions for sentences (like globally ambiguous ones) for which people find it difficult to compute an interpretation online. Fodor (1982) and Kurtzman and MacDonald (1993) discussed the possibility that certain global quantifier scope ambiguities may never be fully resolved by comprehenders, and perhaps global syntactic ambiguities also may not always receive a definitive resolution (see also Ferreira, Ferraro, & Bailey’s, 2002, account of “good-enough” sentence interpretation and brief remarks about strategic effects in reading below).

4.2. Which Statistics?

Languages exhibit many properties that can be counted; some, such as how often verbs follow nouns, seem more relevant than others, such as the frequency distribution for words in the third position in sentences. In a fully specified theory of language comprehension, it would be clear which statistical regularities people encode and use in processing, and why. Clearly, we do not have anything like that kind of theory in hand; we have some evidence about the use of particular statistics that supports the general theoretical framework. As discussed above, in principle it should not be necessary to specify “the statistics that are relevant to language” a priori because that information should fall out of an automatic procedure: a neural network (or similar formalism) that processes language, subject to constraints imposed by the architecture, representations, and input. This procedure also approximates the experience of the child, for whom the relevant statistics are learned rather than pre-specified. Above, we summarized modeling research that represents important progress toward this ideal, and some models (in limited domains) have generated testable predictions. However, there are practical limits on building large-scale models, and analyzing the behavior of a complex dynamical system becomes difficult. These conditions make it difficult to use a computational model as an independent, bottom-up way of determining which statistics are relevant. For a skeptic, the absence of a complete model creates the possibility that the statistical approach is vacuous because it can explain any result. No matter how an experiment turns out, the argument goes, a researcher can find a statistic or combination of statistics that can account for the pattern of results. The approach is therefore not merely unfalsifiable (i.e., able to fit all patterns that do occur); it can also fit patterns of data that never occur.

While it is important to acknowledge the limits of current knowledge, these concerns are not realistic. First, the methodology used in this research does not involve collecting behavioral data and then finding statistics to fit the results. Rather, researchers test hypotheses developed from several sources: linguistic theory; existing empirical findings; close analyses of examples (e.g., data mining a corpus); and other types of theorizing (e.g., about why languages exhibit particular kinds of statistical regularities; see below). Second, the strategy
of tailoring one’s statistical analysis to fit a particular set of data would be self-defeating, because it results in overfitting: the results will not generalize to other data of the same sort (see Seidenberg & Plaut, 2006, for a discussion of this issue with respect to models of word reading). Think the grain of the behavioral data is such that the number of constraints that actually account for detectable variance is relatively small (although not easily accommodated by factorial designs). The theory states that performance is determined by the aggregate effects of all of one’s experiences with language; out of this highly variable set of data, some very strong regularities arise, and these end up being the ones that can account for observable behavior. Finally, the idea that there will always be a statistical or correlation of statistics to fit any data pattern greatly misjudges the extent to which language structure is constrained. There aren’t unlimited degrees of freedom in accounting for the data because there aren’t unlimited degrees of freedom in how a language can be structured. It is true that the number of language statistics that can be calculated is nearly infinite, but most of them are meaningless. The fact that we can calculate language statistics that do not account for data is not a problem if there are other criteria for determining which statistics are relevant.

4.3. Different Models for Different Phenomena?

Above we suggested that implemented models have been important in developing and testing the probabilistic constraints approach. Such models are as yet limited in scope, and many important linguistic phenomena have yet to be addressed. A deeper concern is that every model is different, i.e., different models have been applied to different phenomena. Where is the integrative model that would subsume the broad range of phenomena that have as yet been investigated using many different models?

We are sympathetic to the concerns that are raised by using different models for different phenomena. It would be a problem if the models that explain one phenomenon, studied using one architecture, are incompatible with the principles that explain some other phenomenon, studied using another architecture. Perhaps such models can only succeed when narrowly focused, as might be seen if a more general model were attempted. There probably is no simple way to address this issue, or a simple, preferable alternative. In the early stages of developing this approach, it has been necessary to merely demonstrate that it is sufficient to account for interesting phenomena, given the general climate of skepticism about statistical methods and connectionist models in the study of language. Further progress would be achieved if, as additional models are developed, researchers were able to identify which general computational properties

2 Perhaps a good analogy is the analysis of evolved potentials (Kutas, van Pottel, & Kihlstrom, this volume). This methodology involves gathering many samples of apparently noisy data; every brain wave is different from every other one. Many different aspects of these waves could be measured and counted, and there is no independent theory of how the waves are generated to indicate which elements are important prior to looking. Averaging across many data samples, however, certain regularities emerge (e.g., systematic displacements of the waveform such as P50, N400 and others). Language may exhibit a greater number of regularities, and we also want a better theory of their sources, but the similarities are noteworthy.
are crucial. These principles are more important than the characteristics that differentiate implemented models. This approach has achieved some success in the domain of single word reading. Researchers have identified a small set of critical computational principles, which have been explored in a succession of models. Each model is slightly different than the others (because of advances in understanding network properties or because they focus on different phenomena), but they are governed by the same principles. It is the set of principles and how they apply to a set of phenomena that constitute the explanatory theory, not the properties of individual models.

Achieving this deeper level of understanding language comprehension requires much more research; more models addressing a broader range of phenomena; comparing different architectures with respect to the same phenomena; analyzing models to identify the properties that are critical to achieving human-like performance. This is an ambitious agenda and it is not clear whether there are enough researchers with the sufficient technical skills and interest in the approach to achieve these goals. Moreover, it is not clear whether it is either feasible or desirable to develop a genuinely integrative model of broad scope. As Seidenberg and Plaut (2006) observed,

The concept of a complete, integrative model is a non sequitur, given the nature of the modeling methodology; particularly the need to limit the scope of a model in order (a) to gain interpretable insights from it and (b) to complete a modeling project before the modeler loses interest or dies. The goal of the enterprise, as in the rest of science, is the development of a general theory that abstracts away from details of the phenomena to reveal general, fundamental principles (Putnam, 1973). Each model serves to explore a part of this theory in progress.

We think it’s important to keep in mind that models are tools, not the goal of the theoretical enterprise. The limitations of individual models are tolerable if they yield insights about puzzling phenomena, generate testable hypotheses, and promote theoretical development.

4.4. Where Do Language Statistics Come From?

Within the constraint satisfaction account, a fine-grained characterization of the statistical regularities constraining the interpretation of ambiguities is important to capturing behavioral data. As much of the above discussion suggests, the complexity of the system makes this accounting a nontrivial enterprise. Some insight into the constraints, and a broader account of language performance, may emerge from addressing the question of the origin of the statistical regularities in language. That is, why do languages exhibit certain statistical properties and not others? At least three forces may modulate the statistics of a language. First, some statistical regularities may be shaped by conceptual structures (McRae, Ferretti, & Amyotte, 1997), so that aspects of our nonlinguistic or pre-linguistic thinking constrains the form of utterances. Second, statistics may be shaped by language producers’ sensitivity to limits on our comprehension abilities, so that producers tailor their utterances to those that are more easily understood, in the process creating statistical
regularities in the language. The extent to which speakers are sensitive to listener needs is not fully resolved, but certainly there are at least some clear examples of speakers tailoring speech to their audience, such as the broad differences in the character of child- and adult-directed speech. Finally, some statistics may emerge from the production process itself. MacDonald (1999) and Gennari and MacDonald (2004) argued for this approach, termed the Production-Distribution-Comprehension (PDC) account, which suggests that certain statistical patterns emerge from language producers' needs to maximize fluency during production. For example, speakers appear to adopt syntactic structures for their utterances at least in part to yield an utterance in which more highly "accessible" words are uttered early, where accessibility here refers to a variety of conceptual, lexical, and perhaps articulatory properties that affect the ease of articulating a particular word or phrase (Bock, 1987). MacDonald (1999), Gennari and MacDonald (2004) and Race and MacDonald (2003) have applied this logic to several different comprehension issues and have argued that comprehenders' preferences to interpret ambiguities in favor of one vs. another alternative structure can be linked to the relative frequency of those alternatives in the language, owing to speakers' and writers' syntactic choices during the production process. These choices in turn stem from biases inherent in the production system, such as to place shorter sentence elements (words or phrases) before longer ones, or to place pauses or smaller optional function words before sections of high-production complexity. If this view is on the right track, then an increased understanding of constraint satisfaction in sentence comprehension will emerge from a better grasp of how the production process promotes certain production choices (word orders, word-structure co-occurrences, structure discourse co-occurrences, etc.) and discourages others.

4.5. Where To Go Next?

In presenting this approach, we have already mentioned several important directions for future research. We will close by mentioning two more. First, as detailed in the chapter by Trueswell and Tanenhaus in this volume, researchers are beginning to expand the range of constraints that comprehenders consider by investigating the extent to which comprehenders integrate the visual scene and other aspects of conversational interaction. This work allows an investigation of comprehension of speech, in contrast to the vast majority of studies discussed in this chapter, which have investigated written language. Second, returning to the written language realm, something that would benefit all theoretical perspectives is to increase our understanding of reading data and its relationship to computational accounts of comprehension processes. Researchers from many theoretical perspectives agree that the theorizing and the data are not well matched, in that certain reading patterns are compatible with radically different interpretations of ambiguity resolution processes (e.g., Lewis, 2000; Tanenhaus, 2004; Van Gompel et al., 2001). This situation may be traced to some combination of imprecision in theoretical claims, inability of reading or other dependent measures to resolve fine-grained predictions about time course, insufficient consideration of the possibility that reading and other dependent measures may reflect comprehenders' strategies so that the data may not be a "pure" reflection of the ambiguity resolution processes. That is, we all know that a novel, a newspaper, and a chemistry textbook elicit different reading behaviors, yet there is very
little appreciation among researchers of the degree to which reading strategies might vary with the nature of filler items or comprehension questions in experiments and the extent to which these strategic components could be affecting reading patterns that are attributed to "automatic" sentence processing operations.

Third, the constraint-based approach affords the opportunity to investigate the relationship between acquisition and skilled performance. The focus in adult comprehension has been on timecourse, specifically the speed with which comprehenders can bring constraints to bear on linguistic input, and there has been relatively little discussion of the learning mechanisms by which comprehenders come to possess the relevant constraints. The claim that the learning is inherently statistical invites research into the extent to which there is continuity between acquisition and adult performance and the extent to which a statistical learning account will prove adequate to explain the child's rapid mastery of language. These questions link to an enormous and ongoing research enterprise in child language acquisition, one with its own controversies and struggles to match theory and data. It is therefore an exciting possibility for theoretical development that the studies of the adult state and the acquisition process in the child might be mutually informative and constraining.

REFERENCES


