Computational Psycholinguistics

Matthew W. Crocker[1]

Department of Computational Linguistics and Phonetics, Saarland University,

66041 Saarbrücken, Germany. crocker@coli.uni-sb.de

A draft chapter for the Blackwell *Computational Linguistics and Natural Language Processing Handbook*, edited by Alex Clark, Chris Fox and Shalom Lappin. This draft formatted on 24th June 2009.

1 Introduction

Computational psycholinguistics is concerned with the development of computational models of the cognitive mechanisms and representations that underlie language processing in the mind/brain. As a consequence, computational psycholinguistics shares many of the goals of natural language processing research, including the development of algorithms that can recover the intended meaning of a sentence or utterance on the basis of its spoken or textual realization. Additionally, however, computational psycholinguistics seeks to do this in a manner that reflects how people process language.

Natural language is fundamentally a product of those cognitive processes that are coordinated to support human linguistic communication and interaction. The study of language therefore involves a range of disciplines, including linguistics, philosophy, cognitive psychology, anthropology, and artificial intelligence. Computational psycholinguistics, perhaps more than any other area, epitomizes interdisciplinary linguistic inquiry: the ultimate goal of the enterprise is to implement models which reflect the means by which linguistic information is stored in, and utilized by, the mind and brain. But beyond modeling of the representations, architectures, and mechanisms that underlie linguistic communication, computational psycholinguistics is increasingly concerned with developing *explanatory* accounts, which shed light on why the human language faculty is the way it is. As such, models of human language processing must ultimately seek to be connected with accounts of language evolution and language acquisition.

This chapter presents some of the historically enduring findings from research in computational psycholinguistics, as well as a state of the art overview of current models and their underlying differences and similarities. While

computational models of human language processing have been developed to account for various levels of language processing — from spoken word recognition and lexical access through to sentence production and interpretation — this chapter will place primary emphasis on models of syntactic processing. It will not be surprising that many accounts of human syntactic processing are heavily informed by computational linguistics, specifically natural language parsing. A traditional approach has been to try to identify parsing algorithms which exhibit the range of observed human language processing behaviors, including incremental processing, local and global ambiguity resolution, and parsing complexity (both time and space; see Chapters 2 and 4). Such symbolic approaches have the advantage of being well-understood computationally, transparent with respect to their linguistic basis, and scaleable. An alternative approach has been to develop models using neurally inspired connectionist networks (see Chapter 10), which are able to learn from sufficient experience to language, are robust, and degrade gracefully (Elman, 1990; Plunkett & Marchman, 1996). Purely connectionist approaches often use distributed, rather then symbolic representations, making it difficult to understand precisely what kinds of representations such networks develop. Furthermore, they are typically relatively small scale models, and it has proven difficult to scale their coverage. Some cognitive models of language are in fact best viewed as hybrids, exploiting a mixture of symbolic representations, and connectionist-line computational mechanisms. Most recently, probabilistic approaches have dominated, providing a transparent linguistic basis on the one hand, with an experience-based mechanism on the other.

Before considering the range of approaches, it is important to understand precisely the goals of computational psycholinguistics, and the kinds of data

that inform the develop of models. Furthermore, while many ideas and algorithms have their roots in computational linguistics, we begin by identifying where these two endeavors diverge, and why.

2 Computational models of human language processing

While psycholinguistic theories have traditionally been stated only informally, the development of computational models is increasingly recognized as essential. Specifically, computational models entail the explicit formalization of theories, and also enable prediction of behavior. Implemented models are especially important not only because human language processing is highly complex, involving interaction of diverse linguistic and non-linguistic constraints, but also because it is inherently a dynamic process: People are known to understand, and produce, language incrementally as they read or hear a sentence unfold. This entails that the recovery of meaning happens in real-time, with the interpretation being influenced by a range of linguistic, non-linguistic and contextual sources of information, on the one hand, and also shaping our expectations of what will come next, on the other.

How is computational psycholinguistics different from computational linguistics? In fact, early conceptions of natural language processing explicitly approached language as a cognitive process (Winograd, 1983). Ultimately, however, research is shaped by the specific goals of a particular research community. To understand this more clearly, it can be helpful to distinguish accounts of linguistic *competence* and *performance*. Broadly speaking, a theory of linguistic competence is concerned with characterizing what it means to "know" language, including the kinds syntactic and semantic rules and representations provided by a linguistic theory. A theory of performance, in contrast, characterizes the means by which such knowledge is used *on-line* to recover the meaning for a given sentence, as exemplified by a psychological plausible parsing algorithm.

Consider, for example, one of the classic examples from psycholinguistics, known as the main verb/reduced-relative clause ambiguity Bever (1970):

(1) "*The horse raced past the barn fell*"

For many readers, this sentence seems ungrammatical. The confusion arises because the verb *raced* is initially interpreted as the main verb, leading the parse "*up the garden path*" (Frazier, 1979). Only when the true main verb *fell* is reached can the reader potentially determine that *raced past the barn* should actually have been interpreted as a reduced-relative clause (as in *The horse which was raced past the barn fell*). In this relatively extreme example of a *garden-path* sentence, many readers are unable to recover the correct meaning at all, despite the sentence being perfectly grammatical (cf. *The patient sent the flowers was pleased* which is rather easier, but has the same structure). Thus our linguistic competence offers no explanation for this phenomena, rather it seems necessary to appeal to *how* people recover the meaning, resolving ambiguity as they encounter the sentence incrementally.

Computational linguistics and psycholinguistics have traditionally shared assumptions regarding linguistic competence; both are concerned with developing algorithms which recover a linguistically adequate representation of a sentence as defined by current syntactic and semantic theories. At the level of performance, however, computational linguistics is rarely concerned with issues such as incremental sentence processing and the resolution of *local* ambiguities which are resolved by the end of the sentence. There is rather a greater interest in optimizing the computational properties of parsing algorithms, such as their time and space complexity (see Chapters 2 and 4). Computational psycholinguistics, in contrast, places particular emphasis on the incremental processing behavior of the parser.

As computational linguistics has increasingly shifted its focus towards application domains, the demands of these applications has further divided the computational linguistics and computational psycholinguistics communities. The acknowledged difficulty of computationally solving the *natural language understanding problem*, which in turn relies on a solution of the *artificial intelligence problem*[1], has led to an increased focus in computational linguistics on developing less linguistically ambitious technologies which are scaleable and able to provide useful technologies for particular sub-problems. Robust methods for part-of-speech tagging, named-entity recognition, and shallow parsing (see Chapter 16), for example can contribute to applications ranging from spam-filtering and document classification to information extraction, question answering, and machine translation (see Chapters 20, 24, and 21). For the most part, however, the methods used to perform these tasks have no cognitive basis.

While the research goals of computational linguistics and computational psycholinguistics have diverged since the 1970s, there continues to exist a significant overlap in some the methods that are exploited. An interesting result of the shift towards wide-coverage and robust language processing has been a tremendous emphasis on statistical language processing, and machine learning. As we will see, many of the same underlying methods play a central role in cognitive models as well, with particular overlap coming from research on statistical language modeling (see Chapter 3).

[1] The essence of this argument is that understand language ultimately requires full intelligence — requiring both extensive world knowledge and reasoning abilities — which remain out of reach in the general case (Shapiro, 1992).

2.1 Theories and Models

In developing accounts of human language processing, as with any other cognitive process, it is valuable to distinguish the expression of the theory from a given model which implements the theory. Theories typically relate to a particular aspect of language processing — such as lexical access, parsing, or production — and as such provide incomplete characterizations of language processing in general. Furthermore, theories often provide a relatively high-level characterization of a process, leaving open details about what specific algorithms might be used to realize the theory. Marr (1982), in fact, identifies three levels at which cognitive processes may be described: (1) the *computational* level, which defines *what* is computed, (2) the *algorithmic* level, which specifies *how* computation takes place, and (3) the *implementation* level, which states how the algorithms are actually realized in the neural assemblies and substrates of the brain. In the case of language processing, which is a relatively high-level cognitive function, there have been very few accounts at the third level: we simply have insufficient understanding about how language is processed and represented at the neural level.

There are several reasons for why a distinction of these levels is important. One reason for wishing to state theories at a relatively high level is to emphasize the general properties of the system being described, and ideally some justification of why it is the way it is. Additionally, it is often the case that the relevant empirical data available may not permit a more detailed characterization. That is to say, in building a specific model (at the algorithmic level) of a given theory (stated at the computational level), we are often required to specify details of processing which are underdetermined by the empirical data. While resolving those details is essential to building computational mod-

els that function, we may not wish to ascribe any psychological reality to all aspects of the model. In the event that there is some new piece of empirical evidence which the model incorrectly accounts for, such a distinction is critical: it may be a consequence of the original theory, either falsifying it or entailing some revision to it, or it may simply be a result of some (possibly purely pragmatically-based) decision made in implementing the model, such that only a change at the algorithmic or implementation level, and not the computational level, is needed.

Theories of human language processing can be broadly characterized by the extent to which they assume the mechanisms underlying language processing are *restricted* or *unrestricted* (Pickering *et al.*, 2000a). Restricted accounts begin with the assumption that cognitive processes are resource bound, and that observed processing difficulties in human language processing are a consequence of utilizing or exceeding these resource bounds. In order to explain a number of experimentally observed behaviors, a range of restrictions have been identified which may play a role in characterizing the architecture and mechanisms of the human language processor.

Working Memory: The language processor has limited capacity for storing linguistic representations, and these may be exceeded during the processing of certain grammatical structures, such as center-embeddings: "*The mouse [that the cat [that the dog chased] bit] died.*", in which three noun phrases must be maintained in memory before they can be integrated with their respective verbs (Miller & Isard, 1964; Bever, 1970; Gibson, 1991).

Serial Processing: While there may be many structures that can be associated with a sentence during incremental processing, the human parser

only pursues one structure, rather than several or all of them, so as to minimize space complexity. This predicts that if the sentence is disambiguated as having an alternative structure, some form of *reanalysis* will be necessary and cause processing difficulty (Frazier, 1979).

Modularity: Cognitive processes underlying sentence processing are simplified by restricting their representational and informational domains. This enables on-line syntactic processes to operate independently of more general, complex, and time-consuming cognitive processes such as pragmatics, world knowledge and inference (Fodor, 1983).

Unrestricted accounts, in contrast, typically assume that the processing is not fundamentally constrained, and that people are able to bring diverse informational constraints (i.e. interactive rather than modular) to bear on deciding among possible structures and interpretations (i.e. parallel rather than serial). Such accounts don't deny that cognitive resources are ultimately limited, but do tacitly assume that the architectures and mechanisms for language processing aren't fundamentally shaped by the goal of conserving such resources. Most current models are best viewed as lying somewhere between the two extremes.

2.2 Experimental Data

As noted above, models of human language processing seek not only to model linguistic competence — the ability to relate a sentence or utterance with its intended meaning — but also human linguistic performance. Language processing is best viewed as a dynamic process, in which both the linguistic input and associated processing mechanisms unfold over time. Evidence concerning *how* people process language can be obtained using a variety of methods.

An important aspect of all controlled psycholinguistic experiments, however, is a clear experimental design. Experiments are designed to test a specific hypothesis about language processing, usually as predicted by a particular theoretical proposal or model. As an example, let's consider the matter of serial, incremental processing: the claim that each word is attached into a single connected partial syntactic representation as the sentence is read. This claim makes the prediction that any local ambiguity will be resolved immediately and if that decision later turns out to be wrong, then some processing difficulty will ensue. Consider the following sentences:

(2) a. "*The athlete* [$_{\text{VP}}$ *realized* [$_{\text{NP}}$ *her potential*]]"

b. "*The athlete* [$_{\text{VP}}$ *realized* [$_{\text{S}}$ [$_{\text{NP}}$ *her potential*] [$_{\text{VP}}$ *might make her famous*]]]"

c. "*The athlete* [$_{\text{VP}}$ *realized* [$_{\text{S}}$ [$_{\text{NP}}$ *her exercises*] [$_{\text{VP}}$ *might make her famous*]]]"

In sentence (2a&b), a local ambiguity occurs when we encounter the word "*her*", following the verb "*realized*". While the word "*her*" certainly begins a noun phrase (NP), that NP can either be the direct object of the verb, as in the sentence (2a), or the subject of an embedded sentence, as in (2b). To investigate whether or not people immediately consider the direct object reading, Pickering *et al.* (2000b) compared processing of this ambiguity, manipulating only whether the NP following the verb was a plausible direct object. They argued that if people favor building the direct object reading, this will influence processing complexity in two ways. Firstly, in (2b) they will attach the NP "*her potential*" as the direct object, and then be surprised when they encounter the following VP, which forces them to reanalyzse the object NP as the subject of the embedded clause. For (2c), they should also attempt the direct object attachment, but be surprised because it is implausible, and then assume it begins an embedded clause. In an eye-tracking study,

they found evidence supporting exactly this prediction. Namely, in (2c) people spent longer reading the NP ("*her exercises*") following the verb, than they did reading the NP ("*her potential*") in (2b), suggesting they built a direct object structure only to realize it is implausible. In (2b), however people spent longer reading the *disambiguating* region (the embedded VP) than in (2c), suggesting they had committed to the (plausible) direct object reading, and the needed to revise that analysis.

Since many different factors are known to influence reading times, most psycholinguistic experiments use a design like the one just described above, in which the *difference* in reading times for similar sentences (or regions of the sentence) are compared, and where only the factor which is of interest is varied between the sentences. One simple method which has been used effectively to investigate incremental reading processes is the *self-paced reading* (SPR) paradigm. Using this method, the sentence is presented one word at a time, and the participant must press a key to see the next word. The latency between key presses can then be averaged across both participants and a range of linguistic stimuli to obtain average reading times, which can then be analyzed to determined if there are statistically significant differences in reading times resulting from the experimental manipulation. Another more sophisticated method, eye-tracking, provides an especially rich, real-time window into language processing, with the added advantage of not requiring any additional (possibly unnatural) task. Current eye-tracking technology enables the precise spatial and temporal recording of eye-movements (saccades) and fixations as people read a sentence which is displayed in its entirety on a display. Since people often look back to earlier points in the sentence while reading, several reading-time *measures* can be computed, such as *first-pass* (the amount of

time spent in a region before people move out of the region), or *total-time* (all the time spent reading a region, including looking back at it, etc.) (Rayner, 1998).

When relating a theory or model of language processing to empirical data, it is important to be clear about the exact nature of the relationship that is being assumed, via a *linking hypothesis*. In the example describe above, we implicitly assumed that it was the *surprise* — of either an implausible interpretation, or a subsequent cue that reparsing would be required — that would lead to increased reading time. But there are many characteristics of a computational model that one might argue would be reflected in empirically observable processing complexity. As we'll see below, everything from the frequency of the word which is being processed, to the memory load associated with processing completely unambiguous sentence can be observed in reading times. This is one reason why carefully controlled experiments are so essential, as are clear linking hypothesis that can be used to relate a processing model to some empirical measure.

Reading times offer a robust and well-understood *behavioral* method for establishing processing difficult during sentence comprehension. More recently, however, neuroscientific methods have become increasingly important to informing the development of psycholinguistic theories. This is particularly true of *event-related potentials* (ERPs), which can be observed using electroencephalography (EEG) methods. ERPs reflect brain activity, as measured by electrodes positioned on the scalp, in response to a specific stimulus. Numerous ERP studies have demonstrated the incrementality of language comprehension as revealed by the on-line detection of semantic (e.g., Kutas & Hillyard, 1980, 1983; van Petten & Kutas, 1990) and syntactic (e.g., Osterhout & Holcomb,

1992, 1993; Matzke *et al.*, 2002) violations, indexed broadly by so-called N400 and P600 deflections in scalp activation, respectively. However, while there are several theoretical processing accounts which are derived from such data (Friederici, 2002; Bornkessel & Schlesewsky, 2006), relatively few have led to the development of computational models (but see Crocker *et al.*, in press). For this reason, we will focus here primarily on models based on behavioral findings.

Finally, the *visual world paradigm*, in which participants' eye movements to visually displayed objects are monitored as participants listen to an unfolding utterance, has revealed that people automatically map the unfolding linguistic input onto the objects in their visual environment in real-time (Tanenhaus *et al.*, 1995). Using this method, Allopenna *et al.* (1998) demonstrated not only that increased inspections of visually present objects often occur within 200ms of their mention, but also that such utterance-mediated fixations even reveal sub-lexical processing of the unfolding speech stream. Perhaps of even greater theoretical interest are the findings of Tanenhaus *et al.* (1995), revealing on-line interaction of visual and linguistic information for sentences such as "*Put the apple on the towel in the box*". Not only did listeners rapidly fixate the mentioned objects, but their gaze also suggested the influence of the visual referential context in resolving the structural ambiguity in this sentence (namely, whether *towel* is a modifier of, or the destination for, the *apple*). In fact, this paradigm has also shown that comprehension is not just incremental, but often highly *predictive*: Altmann & Kamide (1999) demonstrated that listeners exploit the selectional restrictions of verbs like *eat*, as revealed by anticipatory looks to edible objects in the scene (before those

objects have been referred to) (see also (e.g. Federmeier, 2007), for related findings from event-related potential studies).

3 Symbolic Models

Evidence that people understand language incrementality is perhaps one of the most ubiquitous findings in experimental research on human sentence processing. The importance of this finding for computational models, is that it places a strong constraint on candidate parsing mechanisms. Not all early computational accounts adhered to the incrementality constraint, however. The *Parsifal* model (Marcus, 1980), for example, proposed a deterministic model of human parsing to account for the observation that people are generally able to understand language in real-time. Parsifal was essentially a bottom-up LR parser, which exploited up to three look-ahead symbols (which could be complex phrases, not just words) to decide upon the next parsing action with certainty. This look-ahead mechanism enabled the parser to avoid making incorrect decisions for most sentences, and Marcus argued that those sentences where the parser failed, were precisely those cases where people also had substantial difficulty.

There are, however several criticisms that can be leveled at Parsifal. Not only is the parser highly non-incremental, with the capacity to leave large amounts of the input on the stack, it also offers only a binary account of processing difficult: easy versus impossible. Experimental research has shown, however, that some kinds of erroneous parsing decisions are much easier to recover from than others (see Sturt *et al.* (1999) for a direct comparison of two such cases). The *licensing-structure* parser (Abney, 1989) responded to these criticisms by adapting a shift-reduce parsing architecture of Pereira (1985) to operate more incrementally. Since look-ahead must be excluded in order to maintain incrementality, the parser often faces nondeterminism during processing. For these cases, Abney proposes several preference strategies which

are intended to reflect parsing principles motivated by human behaviour such as *Right Association* (Kimball, 1973) (attach incoming material low on the right frontier of the current parse) and *Theta Attachment* (Pritchett, 1992) (attach constituents so as to maximize thematic role assignment, see Section 3.1). Abney additionally addressed the issue of *backtracking*, in case parsing fails and an alternative parse needs to be found. The licensing-structure parser, however, is still not strictly incremental, with some parse operations delaying the attachment of words and constituents. A further criticism, which applies to the accounts proposed by Marcus, Abney and Pritchett, is their strong reliance on verb information to determine the parsers actions. While this approach works reasonably for languages like English, it is problematic for explaining parsing of verb-final languages like Japanese, Turkish, and many others.

Resnik (1992), reconsiders the role of space, or memory, utilization as a criteria for selecting psychologically plausible parsing algorithms. As noted above, embedding structures reveal an interesting property of human sentence processing, illustrated by Resnik's following examples (brackets indicate embedded clauses):

(3) a. "[[[*John's*] *brother's*] *cat*] *despises rats*" EASY
    b. "*This is* [*the dog that chased* [*the cat that bit* [*the rat that ate the cheese*]]]" EASY
    c. "[*The rat* [*that the cat* [*that the dog chased*] *bit*] *ate the cheese*]" HARD

While people typically find left embeddings (3a) and right-embeddings (3b) relatively unproblematic, center-embeddings (3c) are often judged as difficult, if not completely ungrammatical (though one can quite clearly demonstrate that they violate no rules of grammar). Building on work previous by Abney & Johnson (1991) and Johnson-Laird (1983), Resnik (1992) demonstrates that

neither strictly top-down (LL) nor bottom-up (LR) parsers can explain this observation. Top-down parsing predicts only right embeddings to be easy, while bottom-up predicts only left embeddings to be easy. Further, Resnik demonstrates that the standard version of a left-corner (LC) parser, which combines top-down and bottom-up parsing is no different than the bottom-up parser with regard to stack complexity. However, an *arc-eager* variant of the LC parser — in which nodes that are predicted bottom-up can be immediately *composed* with nodes that are predicted top-down — models the human performance correctly: left and right embeddings have constant complexity, while center embedding complexity increases linearly with the number of embeddings. A further advantage of the arc-eager LC parser is that it is incremental for all but a few sentence structures (see Crocker (1999) for more detailed discussion of parsing mechanisms).

3.1 Ambiguity resolution

A central element of any model of sentence processing concerns how it deals with lexical and syntactic ambiguity: namely how do we decide which representation to assign to the current input. The assumption of incremental processing further entails that decisions regarding which structure to pursue must be made as each word is processed. Once simple solution to this issue is to propose a parallel model of processing, in which all possible syntactic analyses are pursued simultaneously during processing. Such a solution has traditionally been discarded for two reasons: Firstly, for a large-scale grammar and lexicon, hundreds of analyses may be possible at any point during parsing — indeed, for grammars with left-recursion, there may in fact be an unbounded number of parses possible — and would arguably exceed cognitively plausible memory limitations. One solution to this is to assume *bounded*

parallelism, in which only a limited subset of parses is considered. Secondly, even if one assumes parsing is (possibly bounded) parallel, there is strong evidence that only one interpretation is consciously considered, otherwise we would never expect to observe the kinds of garden-path sentence discussed in Section 2. Thus regardless of whether incremental processing is serial or parallel, any model requires an account of which parse is to be preferred.

There have been many proposals to explain such ambiguity preferences in the psycholinguistic literature. Frazier (1979), building on previous work by Kimball (1973), proposed the following two general principles:

Minimal Attachment (MA): Attach incoming material into the phrase marker being constructed using the fewest nodes consistent with the well-formedness rules of the language

Late Closure (LC): When possible, attach incoming material into the clause or phrase currently being parsed.

Recall example (2), above. When the noun phrase "*her potential*" is encountered, it can either be attached directly as the object of "*realized*" (2a), or as the subject of the embedded clause (2b). The latter structure, however, requires an additional node in the parse tree, namely an S node intervening between the verb and the noun phrase. Thus MA correctly predicts the human preference to interpret the noun phrase as a direct object, until syntactic or semantic information disambiguates to the contrary.

While these parsing principles dominated the sentence processing for some time, they have been criticized on several grounds. Firstly, as noted by Abney (1989) and Pritchett (1992), MA is highly sensitive to the precise syntactic analysis assigned by the grammar. The adoption of binary branching

structures in many modern syntactic theories means that MA fails to differentiate between a number of ambiguities (including the one in Figure 2, discussed below). In response to this, several theories proposed a shift away from MA towards what Pritchett (1992) dubbed *Theta Attachment* (see also Abney (1989); Crocker (1996) for related proposals). Theta Attachment states that the parser should attempt to maximally satisfy verb argument relations whenever possible, and thus prioritize the parsing of phrases into such argument positions, where they will receive a semantic, or *thematic*, role from the verb Fillmore (1968). Returning to sentence (2a), Theta Attachment asserts that attaching the noun phrase "*her potential ...*" as a direct object is preferred because not only is the verb able to assign a thematic role (THEME) to the noun phrase, but the noun phrase also receives a thematic role at that point in processing. If the noun phrase were attached as the embedded subject, as in (2a), it would temporarily have no role assigned to it (until the embedded predicate "*might make ...*" is processed. Thus the latter option is dispreferred.

The above approaches are typically associated with modular processing accounts (Fodor, 1983), since they emphasize the role of purely syntactic decision strategies for parsing and disambiguation. Serial parsing is also assumed, namely that the human language processor only constructs one parse — backtracking or reanalyzing the sentence if that parse turns out to be incorrect. For these reasons, such models of processing are typically viewed as *restricted* accounts, since they fundamentally assume a processing architecture which is limited by the kinds of information it has access to (i.e. syntactic), and the memory resources available for parsing.

While there is a considerable body of experimental evidence supporting the importance of such syntactic strategies, there is also evidence suggesting that people are nonetheless able to draw upon a large repertoire of relevant constraints in resolving ambiguity, such as specific lexical biases, and semantic plausibility (Gibson & Pearlmutter, 1998). The general claim of such *interactive constraint-based* approaches is that parsing is not a serial process influenced solely by syntactic strategies, but rather that "*multiple alternatives are at least partially available, and that ambiguity resolution is accomplished by the use of correlated constraints from other domains*" (Trueswell & Tanenhaus, 1994). While one might envisage such a model in symbolic terms, they typically rely on the use of probabilistic constraints, and are better viewed as *hybrid* models, which will will discuss in Section 6.

3.2 Working memory

The above discussion of the left-corner parser above might lead one to believe that center-embeddings are the only unambiguous syntactic structures which cause processing difficulty. Gibson (1991), however, argues that processing complexity arising from working memory demands can also explain ambiguity resolution preferences. Building on Pritchett (1992)'s *Theta Attachment* strategy, Gibson attributes a *cost* to the parser's need to maintain thematic role-assignments and role-fillers in memory. He argues that such a working-memory metric can be used not only to explain increased processing complexity for structures with locally high memory demands, but also that his metric can be used to rank candidate parsers in the face of local ambiguity. That is, the parser will generally prefer interpretations which have lower cost with respect to unfulfilled role-relations, thus predicting disambiguation behavior in a manner similar to (Pritchett, 1992). Gibson (1998)'s *dependency*

locality theory refines this approach further, by taking into account the distance between role-assigners and role-recipients (see also (Gibson, 2003) for an overview). Lewis *et al.* (2006) propose an account of parsing which draws on a number of general observations concerning the dynamics of memory retrieval that have been established across cognitive domains. These principles have also been implemented within general of cognitive architecture ACT-R (Anderson *et al.*, 2004), enabling Lewis and colleagues to provide an independently motivated proposal regarding the role of working memory in sentence processing.

4 Probabilistic Models

The symbolic accounts outlined above offer insight into both how hierarchical sentence structure and meaning are recovered incrementally, and when processing of such sentences may be difficult as a consequence of either working memory limitations or the need to reanalyze the sentence if the parser has followed a garden-path. A variety of empirical results, however, suggest that such symbolic, modular and serial processing mechanisms may not scale sufficiently to account for human linguistic performance in general (Crocker, 2005). Firstly, serial backtracking parsers are known to be extremely inefficient as grammars are scaled up to provide realistic linguistic coverage. In addition, such models accord no role to linguistic experience despite a wealth of experimental findings indicating that frequency information plays a central role in determine the preferred part of speech, meaning, and subcategorization frame for a given word. Finally, while cognitive resources like working memory undoubtedly constrain language processing, and provide an index of certain kinds of processing complexity, it has been argued that people are in general able to understand most language effectively and without conscious effort. Indeed, one of the most challenging tasks facing computational psycholinguistics, is to explain how people are able deal with the complexity and pervasive ambiguity of natural language so accurately and in real-time: what Crocker (2005) dubs *the performance paradox*.

Probabilistic approaches offer a natural means to address the above issues. Not only do they provide a means to develop *experience-based* models, which can exploit the kinds of frequency information that people have been shown to use, but probabilistic methods have also proven extremely successful for developing wide-coverage models of language processing (see Chapters 3, 4

and 14). Perhaps more fundamentally, probabilistic methods invite us to view language processing less in terms of the difficulties people exhibit on some kinds of constructions, and instead emphasize the remarkable performance that people exhibit in understanding language in general. Chater *et al.* (1998) explicitly argue that human language processing may be fruitfully viewed as a *rational* process, in Anderson (1990)'s sense of the term. If one views language understanding as a rational cognitive process one can begin by first identifying the *goal* of that process — e.g. to find the correct interpretation of a sentence — and then reason about the function that best achieves that goal and accounts for observed behavior. One obvious rational analysis of parsing is to assume that the parser chooses operations so as to maximize the likelihood of finding the intended global interpretation of the sentence, taking into account known cognitive and environmental limitations. Given the overwhelming evidence that people process language incrementally, we can plausibly define the function that is implemented by the human language processor as follows:

$$\hat{t_i} = \underset{t_i}{\operatorname{argmax}}\ P_i(t_i|w_{1\ldots i}, K), \forall t_i \in T_i \tag{1}$$

This states that as each word w_i is processed, the preferred analysis of the sentence initial substring $w_1 \ldots w_i$, $\hat{t_i}$, corresponds to the analysis t_i — in the set of possible analyses T_i that span the sentence up to and including w_i — that has the highest likelihood given the words of the sentence, and our general knowledge K.[2] Crucially, this equation provides only a high-level characterization of how people process language, namely at Marr's *compu-*

[2] I deliberately use the term *analyses* to abstract away from what particular linguistic representation — lexical, syntactic, semantic, etc. — we might be interested in.

tational level, which we will refer to as the *Likelihood Hypothesis*. It leaves aside many crucial issues concerning *how* the analyses are constructed and their probabilities estimated. In principle, the likelihood of a particular analysis of a sentence might reflect not only our accumulated linguistic experience as it relates to the current input, but also the current context and our general knowledge K. But just as statistical language processing techniques have vastly simplified the kind of information used to condition the probabilities, it may be reasonable to assume that people similarly approximate probabilities, at least during initial processing of the input. In the following sections we review several proposals that can be viewed as instances of the Likelihood Hypothesis.

4.1 Lexical Processing

Much of the ambiguity that occurs in syntactic processing in fact derives from ambiguity at the lexical level (MacDonald *et al.*, 1994). Furthermore, it is precisely at the lexical level that frequency effects have been most robustly observed: high frequency words are processed more quickly than low frequency ones (Grosjean, 1980); words are preferentially understood as having their most likely part of speech (Trueswell, 1996; Crocker & Corley, 2002); verbs subcategorization preferences rapidly influence parsing decisions (Ford *et al.*, 1977; Garnsey *et al.*, 1997; Trueswell *et al.*, 1993); and semantically ambiguous words are preferably associated with their more frequent sense (Duffy *et al.*, 1988). These findings all suggest that a likelihood based resolution of lexical ambiguity will substantially reduce parsing ambiguity, and assist in guiding the parser toward the most likely parse in a manner that reflects human behavior.

Based on this rationale, Corley & Crocker (2000) propose a broad coverage model of lexical category disambiguation as a means for substantially constraining the preferred syntactic analysis. Their approach uses a bi-gram model to incrementally determine the most probable assignment of part of speech tags, $\hat{t_0} \ldots \hat{t_i}$, for the (sub-)string of input words $w_0 \ldots w_i$, as follows:

$$\hat{t_0} \ldots \hat{t_i} = \underset{t_0 \ldots t_i}{\operatorname{argmax}} P(t_0 \ldots t_i, w_0 \ldots w_i) \approx \prod_{j=1}^{i} P(w_j|t_j)P(t_j|t_{j-1}) \qquad (2)$$

The bi-gram model results in the use of both the unigram likelihood of word w_j given a possible part of speech t_j, $P(w_j|t_j)$, as well as the context as captured by the immediately preceding part of speech tag $P(t_j|t_{j-1})$. The likelihood for a particular sequence of parts-of-speech, spanning from w_0 to w_i, is the product of this value as computed for each word in the string. In order to efficiently determine the most likely part of speech sequence as the sentence is processed, the Viterbi algorithm is used (Viterbi, 1967).

(4) a. "*The warehouse prices are cheaper than the rest*"

b. "*The warehouse makes are cheaper than the rest*"

This model capitalizes on the insight that many syntactic ambiguities have a lexical basis, as in (4). These sentences are ambiguous between a reading in which "*prices*" (4a) or "*makes*" (4b) serves as either the main verb or part of a compound noun. Once trained on a large corpus, the model predicts the most likely part of speech for "*prices*", correctly accounting for the fact that people preferentially interpret "*prices*" as a noun, but "*makes*" as verb (Frazier & Rayner, 1987; MacDonald, 1993). In the latter case, a difficultly in processing is observed once the sentence disambiguates "*makes*" as a noun (Crocker & Corley, 2002). The model similarly accounts for the

finding that categorially ambiguous word like "*that*" are resolved by their preceding context: In sentence initial position, "*that*" is more likely to be a determiner, while post-verbally, it is more likely to be a complementiser (Juliano & Tanenhaus, 1993).

Interestingly, the use of the Viterbi algorithm to determine the most likely sequence incrementally, predicts that reanalysis may occur when the most probable part of speech sequence at a given point requires revising a preceding part of speech assigment. This behavior in the model finds support from a study by (Macdonald, 1994) showing that reduced-relative clause constructions, like those illustrated in (5) were rendered easier to process when the word following the ambiguous verb (simple past vs. participle) made the participle reading more likely.

(5) a. "*The sleek greyhound admired at the track won four trophies*"

b. "*The sleek greyhound raced at the track won four trophies*"

Since "*admired*" (5a) is transitive, the fact that is it not followed by a noun phrase is a clear cue that it's part of speech should be past-participle, and parse inside the relative clause. For "*raced*" (5b), however, which is preferentially intransitive, the preposition "*at*" provides no such cue for rapid reanalysis, resulting in a garden path when the main verb "*won*" is reached.

Importantly, however, the model not only accounts for a range of disambiguation preferences rooted in lexical category ambiguity, it also offers an explanation for why, in general, people are highly accurate in resolving such ambiguities. It's also worthwhile to distinguish between various aspect of this account in terms of Marr's three levels. Equation (2) provides the *computational* theory, the likelihood function defining the goal of the process, and it's *algorithmic* instantiation in terms of the bi-gram model and the Viterbi

algorithm. This highlights the point that one might change the algorithmic level — e.g. by using a tri-gram model should there be empirical evidence to support this — without in any way changing the computational theory. The *implementation* level is not provided, since this would entail a characterization of how the bi-gram model is processed in the brain and how probabilities are estimated over the course of our linguistic experience (which we simply approximate using corpus frequencies).

4.2 Syntactic Processing

While lexical disambiguation is an important part of sentence processing, and goes a considerable way towards resolving many structural ambiguities, Corley & Crocker (2000)'s model is clearly not a full model of syntactic processing. Indeed, Mitchell *et al.* (1995) have taken the stronger view that the human parser not only makes use of lexical frequencies, but also keeps track of *structural* frequencies. Evidence from relative clause attachment ambiguity (see Figure 4.2) has been taken to support an experience-based treatment of structural disambiguation. Such constructions are interesting because they do not hinge on lexical preferences. When reading sentences containing the ambiguity in Figure 4.2, English comprehenders appear to follow Frazier's Late Closure strategy, demonstrating a preference for low-attachment (where "*the actress*" is modified by the RC "*who . . .*"). Spanish readers, in contrast, when presented with equivalent Spanish sentences, prefer high-attachment (where the RC concerns "*the servant*") (Cuetos & Mitchell, 1988). This finding provided evidence against the universality of Frazier's Late Closure strategy (Section 3.1), leading Mitchell *et al.* (1995) to propose the *Tuning Hypothesis*, which asserts that the human parser deals with ambiguity by initially selecting the syntactic analysis that has worked most frequently in the past (Brysbaert & Mitchell,

1996). Later experiments further tested the hypothesis, examining school children's preferences before and after a period of two weeks in which exposure to high or low examples was increased. The findings confirmed that even this brief period of variation in *experience* influenced the attachment preferences as predicted (Cuetos *et al.*, 1996).

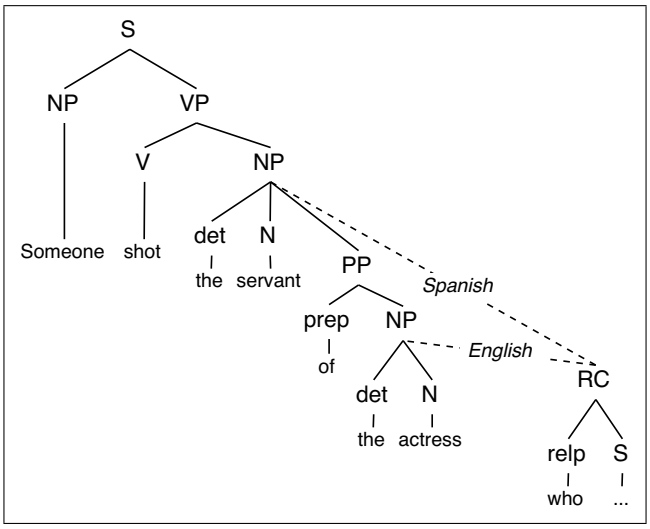


Figure 1. Relative clause attachment ambiguity

Models of human syntactic processing have increasingly exploited probabilistic grammar formalisms, such as Probabilistic Context Free Grammars (PCFGs) to provide a uniform probabilistic treatment of lexical and syntactic processing and disambiguation (For PCFGs, see Manning & Schütze 1999, as well as Chapters 3, 4 and 14). PCFGs augment standard context free grammars by annotating grammar rules with rule probabilities. A rule probability expresses the likelihood of the left-hand side of the rule expanding to its righthand side. As an example, consider the rule VP → V NP in Fig. 2a. This rule says that a verb phrase expands to a verb followed by a noun phrase with

a probability of 0.7. In a PCFG, the probabilities of all rules with the same lefthand side must sum to one:

$$\forall i \sum_{j} P(N^i \rightarrow \zeta^j) = 1 \tag{3}$$

where $P(N^i \rightarrow \zeta^j)$ is the probability of a rule with the lefthand side N^i and the righthand side ζ^j. For example, in Fig. 2a the two rules VP $\rightarrow$ V NP and VP $\rightarrow$ VP PP share the same lefthand side (VP), so their probabilities sum to one. The probability of a parse tree generated by a PCFG is computed as the product of its the rule probabilities:

$$P(t) = \prod_{(N \rightarrow \zeta) \in R} P(N \rightarrow \zeta) \tag{4}$$

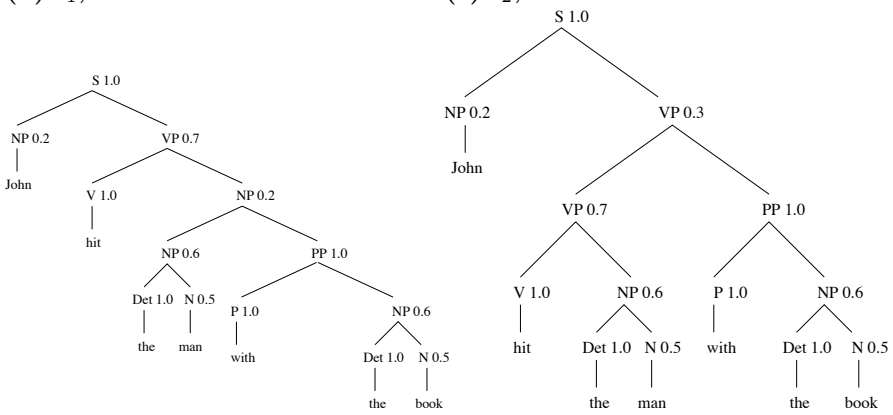
where R is the set of all rules applied in generating the parse tree t. While rule probabilities are in theory derived during the course of a persons linguistic experience, most models rely on standard techniques for estimating probabilities such as *maximum likelihood estimation* — a *supervised* learning algorithm which calculates the probability of a rule based on the number of times it occurs in a parsed training corpus. An alternative, *unsupervised* method is the expectation maximization (EM) algorithm, which uses an unparsed training corpus to estimate a set of rule probabilities that makes the sentences in the corpus maximally likely (Baum, 1972) (see also Chapter 9).

Just as lexical frequency may determine the ease with which words are retrieved from the lexicon, and the preferred morphological, syntactic and semantic interpretations we associate with them, Jurafsky (1996) argues that the probability of a grammar rule corresponds to how easily that rule can be accessed by the human sentence processor during parsing. The consequence of

(a)

S → NP VP	1.0	NP → Det NP	0.6	V → hit	1.0
PP → P NP	1.0	NP → NP PP	0.2	N → man	0.5
VP → V NP	0.7	NP → John	0.2	N → book	0.5
VP → VP PP	0.3	P → with	1.0	Det → the	1.0

(b) t_1**;** **(c)** t_2**;**



$$P(t_1) = 1.0 \times 0.2 \times 0.7 \times 1.0 \times 0.2 \times 0.6 \times 1.0 \times 1.0 \times 0.5 \times 1.0 \times 0.6 \times 1.0 \times 0.5 = 0.00252$$

$$P(t_2) = 1.0 \times 0.2 \times 0.3 \times 0.7 \times 1.0 \times 1.0 \times 0.6 \times 1.0 \times 0.6 \times 1.0 \times 0.5 \times 1.0 \times 0.5 = 0.00378$$

Figure 2. An example for the parse trees generated by a probabilistic context free grammar (PCFG), (adapted from Crocker & Keller, 2006). (a) The rules of a simple PCFG with associated rule application probabilities. The two parse trees, (b) and (c), generated by the PCFG in (a) for the sentence "*John hit the man with the book*", with the respective parse probabilities, $P(t_1)$ and $P(t_2)$, calculated below.

this claim is that structures with greater overall probability should be easier to construct, and therefore preferred in cases of ambiguity. The PCFG in Fig. 2a generates two parses for the the sentence "*John hit the man with the book*". The first parse t_1 attaches the prepositional phrase "*with the book*" to the noun phrase (low attachment) with a total probability of 0.00252 (see Fig. 2b). The alternative parse t_2, with the prepositional phrase attached to the verb phrase (high attachment) is assigned a probability of 0.00378 (see

Fig. 2c). Under the assumption that the probability of a parse determines processing ease, the grammar will predict that t_2 (high attachment) will be generally preferred to t_1, as it has a higher probability.

In applying PCFGs to the problem of human sentence processing, Jurafsky (1996) makes two important observations. First he assumes that parsing, and the computation of parse probabilities, takes place incrementally. The consequence is that the parse faces a local ambiguity as soon as it hears the fragment "*John hit the man with ...*" and must decide which of the two possible structures is to be preferred. This entails that the parser is able to compute prefix probabilities for sentence initial substrings, as the basis for comparing alternative (partial) parses (Stolcke, 1995). For the example in Fig. 2, it should be clear that the preference for t_2 would be predicted even before the final NP is processed, since the probability of that NP is the same for both structures.

The second major contribution of Jurafsky (1996)'s approach is the proposal to combine structural probabilities generated by a probabilistic context free grammar with probabilistic preferences of individual lexical items, using Bayes' Rule. The model therefore integrates lexical and syntactic probabilities within a single mathematically founded probabilistic framework. As an example consider the sentences in (6), which have a similar syntactic ambiguity to that outlined in Fig. 2.

(6) a. "*The women discuss the dogs on the beach*"

b. "*The women keep the dogs on the beach*"

The intuition when one reads these sentences is that low-attachment of the PP "*on the beach*" to the NP "*the dogs*" is preferred for (6a), while high-attachment to the verb is preferred for (6b). A standard PCFG model, how-

ever, will always prefer one of these (in our example PCFG, high-attachment). Following Ford *et al.* (1977), Jurafsky argues that we must also take into account the specific subcategorization preferences of the verb (Table 1), in addition to the structural probabilities of the PCFG.

Table 1. Conditional probability of a verb frame given a particular verb, as estimated using the Penn Treebank.

Verb	Frame	P(Frame|Verb)
discuss	<NP PP>	.24
	<NP>	.76
keep	<NP PP>	.81
	<NP>	.19

Jurafsky's model computes the probabilities of these two readings based on two sources of information: the overall structural probability of the high-attachment reading and the low-attachment reading, and the lexical probability of the verb occurring with an <NP PP> or an <NP> frame. The structural probability of a reading is independent of the particular verb involved; the frame probability, however, varies with the verb. This predicts that in some cases lexical probabilities can override the general structural probabilities derived from the PCFG. If we combine the frame probabilities from Table 1 with the parse probabilities determined with the PCFG in Fig. 2, we can see that high-attachment preference is maintained for "*keep*", but low-attachment becomes more likely for "*discuss*".

Strictly speaking, Jurafsky's model does not aim to recover the single most likely parse during processing, as suggested in Equation 1. Rather he argues for a bounded parallel model, which pursues the most probable parses and prunes those parses whose probability is less than $\frac{1}{5}$ the probability of

the mostly likely parse. Strong garden paths are predicted if the ultimately correct syntactic analysis is one which has been pruned during parsing.

4.3 Wide coverage models

Jurafsky (1996) outlines how his framework can be used to explain a variety of ambiguity phenomena, including cases like (4) and (5) discussed above. However, it might be criticized for its limited coverage, i.e., for the fact that it uses only a small lexicon and grammar, manually designed to account for a handful of example sentences. Given that broad coverage parsers are available that compute a syntactic structure for arbitrary corpus sentences, it is important that we demand more substantial coverage from our psycholinguistic models to insure they are not *over fitting* to a small number of garden path phenomena.

Crocker & Brants (2000) present the first attempt to develop a truly wide-coverage model, based on the incremental probabilistic parsing proposals of Jurafsky (1996). Their approach combines on the wide-coverage psycholinguistic bi-gram model of Corley & Crocker (2000) with the efficient statistical parsing methods of Brants (1999). The resulting *incremental cascaded Markov model* has broad coverage, relatively good parse accuracy in general, while also accounting for a range of experimental findings concerning lexical category and syntactic ambiguities. For practical reasons, Crocker & Brants (2000) don't include detailed subcategorization preferences for verbs, but rather limit this to transitivity, which is encoded as part of a each verbs part of speech. Adopting a parallel parsing approach not unlike that of Jurafsky, Crocker & Brants (2000) also argue that *re-ranking* of parses, not just pruning of the correct parse, is a predictor of human parsing complexity.

This research demonstrates that when such models are trained on large corpora, they are indeed able to account for human disambiguation behavior such as that discussed by Jurafsky (1996). In related work, Brants & Crocker (2000) also demonstrate that broad coverage probabilistic models maintain high overall accuracy even under strict memory and incremental processing restrictions. This is important to support the claim that rational models maintain their *near optimality* even when subject to such cognitively motivated constraints.

4.4 Information Theoretic Models

The probabilistic parsing proposals of Jurafsky (1996) and Crocker & Brants (2000) provide relatively coarse-grained predictions concerning human processing difficulty, based on whether or not the ultimately correct parse was assigned a relatively low probability (or pruned entirely), and must be re-ranked (or even re-parsed). Drawing on concepts developed in the statistical language modeling literature (see Chapter 3), Hale (2001) proposes a more general linking hypothesis between incremental probabilistic processing and processing complexity. Specifically, Hale suggests that the cognitive effort associated with processing the next word, w_i, of a sentence will be proportional to its *surprisal*. Surprisal is measured as the negative log probability of a word, such that surprising (unlikely) words contribute greater information than words that are likely, or expected, given the prefix of the sentence, $w_1 \ldots w_{i-1}$.

$$\textit{Effort} \propto -\log P(w_i | w_1 \ldots w_{i-1}, \text{Context}) \approx -\log \frac{P(T_i)}{P(T_{i-1})} \tag{5}$$

The notion of information, here, derives from *information theory* (Shannon, 1948), where highly likely or predictable words are viewed has providing little information, while unexpected words provide more. While in principle, all our knowledge about the words $w_1 \ldots w_{i-1}$, linguistics constraints, and non-linguistic *Context* will determine the probability of w_i, Hale assumes the probability can be reasonably approximated by a PCFG. Specifically, he proposes that the probability of a given (sub-)string $w_1 \ldots w_i$ is T_i, which is the sum of all possible parses t_i for the prefix string (Equation 5). Thought of in this way, surprisal at word w_i will be proportional the summed probability of all parses which are *disconfirmed* by the transition from word w_{i-1} to word w_i.

Hale (2001)'s theory thus assumes full parallelism, and can be thought of as associating cognitive processing effort with the sum of *all* disambiguation that is done during parsing. This contrasts with standard accounts in which it is only disconfirmation of the *preferred* interpretation which is assumed to cause processing difficulty. While the assumption of full parallelism raises some concerns regarding cognitive plausibility, Hale's model is able to account for a range of garden-path phenomena as well as processing complexity in unambiguous constructions, such as the dispreferred status of object versus subject relative clauses. In recent work, Levy (2008) refines and extends Hale (2001)'s approach in several respects, improving the mathematical properties of the *surprisal theory* while also extending the empirical coverage of the general approach. Hale (2003) proposes another variant on this approach, the *entropy reduction hypothesis*, in which cognitive effort is linked to a slightly different measure, namely the reduction in uncertainty about the rest of the sentence.

4.5 Probabilistic Semantics

One major limitation of cognitive models of sentence processing is their emphasis on syntactic aspects of processing. This was arguably justified to some extent during the 1980s, when modular theories of language, and cognition in general, prevailed. Since then, however, a wealth of empirical results have shown that semantics and plausibility do not only influence our final interpretation of a sentence, but that such information rapidly informs on-line incremental comprehension. In the case of the probabilistic parsing models discussed above, probabilities are conditioned purely on syntactic and limited lexical frequencies. For primarily practical reasons, a range of independence assumptions are made. Our PCFG above, for example, will assign exactly the same probability to the sentences "*John hit the man with the book*" and "*John hit the book with the man*", since exactly the same rules of grammar are used in deriving the possible parse trees. Yet clearly the latter is semantically implausible, regardless of how it is parsed, and therefore should be assigned a lower probability.

In experimental psycholinguistics, the on-line influence of semantic plausibility has been investigated by varying the argument of a particular verb-argument-relation triple, often called *thematic fit*. McRae *et al.* (1998) investigated the influence of thematic fit information on the processing of the main-clause/reduced-relative clause(MC/RR) ambiguity as illustrated in the sentences below.

(7) a. "*The pirate terrorized by his captors was freed quickly*"

b. "*The victim terrorized by his captors was freed quickly*"

During incremental processing of sentences like (7a), the prefix "*The pirate terrorized . . .*" is ambiguous between the more frequent main-clause continu-

ation (e.g., as in "*The pirate terrorized the Seven Seas*") and a less frequent reduced-relative continuation as shown in (7), where "*terrorized*" heads a relative clause that modifies "*pirate*". The subsequent *by*-phrase provides strong evidence for the reduced-relative reading, signaling the absence of a direct object which would otherwise be required if "*terrorized*" were in simple past-tense, and suggests it is more likely a past participle. Finally the main verb region "*was freed*" completely disambiguates the sentence.

Evidence from reading-time experiments has shown that readers initially have a strong preference for the main-clause interpretation over the reduced relative, but that this preference can be modulated by other factors (e.g. Rayner *et al.*, 1983; Trueswell, 1996; Crain & Steedman, 1985). McRae *et al.* (1998), in particular, showed that good thematic fit of the first NP as an object of the verb in the case of *victim* in (7b) allowed readers to partially overcome the main-clause preference and more easily adopt the dispreferred reduced-relative interpretation, which makes the first NP the object of the verb (as opposed to the main-clause reading, where it is a subject). Reading time effects, both on the ambiguous verb and in the disambiguating region, suggest that the thematic fit of the first NP and the verb rapidly influences the human sentence processor's preference for the two candidate structures.

Narayanan & Jurafsky (1998) outline how Bayesian belief networks can be used to combine a variety of lexical, syntactic and semantic constraints. The central idea is that we can construct a belief network which integrates multiple probabilistic sources of evidence, including: structural probabilities determined by the PCFG; subcategorization preferences as motivated by Jurafsky (1996); verb tense probabilities; thematic fit preferences; and so on. The central problem with this framework is that while extremely powerful and

flexible, there is at present no general method for parsing and constructing such Bayesian belief networks automatically. Rather, the networks must be constructed by hand for each possible structured to be modelled. We therefore leave aside a detailed discussion of this approach, while emphasizing that it may provide a valuable framework for modeling specific kinds of probabilistic constraints (see Jurafsky (2003) for detailed discussion).

In recent work, Pado *et al.* (2009) extend standard probabilistic grammar-based accounts of syntactic processing with a model of human thematic plausibility. The model is able to account for syntactic *and* semantic effects in human sentence processing, while retaining the main advantages of probabilistic grammar based models, namely their ability naturally to account for frequency effects and their wide coverage of syntactic phenomena and unseen input.

The probabilistic formulation of the semantic model equates the plausibility of a verb-argument-role triple with the probability of that thematic role co-occurring with the verb-argument pair — e.g. *terrorized-victim-*AGENT. The semantic model (Equation 6) estimates the plausibility of a verb-role-argument triple as the joint probability of five variables: These are, apart from the identity of the verb v, argument a and thematic role r, the verb's sense s and the grammatical function gf of the argument. The verb's sense is relevant because it determines the set of applicable thematic roles, while the grammatical function linking verb and argument (e.g., *syntactic subject* or *syntactic object*) carries information about the thematic role intended by the speaker.

$$Plausibility_{v,r,a} = P(v, s, gf, r, a) \tag{6}$$

This type of *generative model* can predict the most likely instantiation for missing input or output values, allowing it to naturally solve its dual task of identifying the correct role that links a given verb and argument, and making a plausibility prediction for the triple. It predicts the preferred thematic role for a verb-argument pair, $\hat{r}_{v,a}$, by generating the most probable instantiation for the role, as shown in Equation (7).

$$\hat{r}_{v,a} = \underset{r}{argmax} \;\; P(v, s, gf, r, a) \tag{7}$$

The semantic model is to a large extent derived automatically from training data: clusters of semantically similar noun and verbs are used to reduce the number of *unseen* triples in the semantically annotated FrameNet corpus (Fillmore *et al.*, 2003). The advantage of this approach is that it eliminates the needs to obtain plausibility estimates experimentally (McRae *et al.*, 1998).

In addition to demonstrating that the semantic model reliably predicts a range of plausibility judgement data, Pado *et al.* (2009) integrate the model into a broad-coverage sentence processing architecture. The so-called SynSem-Integration model, shown in Fig. 3, combines a probabilistic parser, in the tradition of Jurafsky (1996) and Crocker & Brants (2000), with the semantic model described above. The syntax model, based on Roark (2001)'s top-down probabilistic parser, incrementally computes all possible analyses of the input and their probabilities. The semantic model evaluates the resulting structures with respect to the plausibility of the verb-argument pairs they contain. Both models simultaneously rank the candidate structures: The syntax model ranks them by parse probability, and the semantic model by the plausibility of the verb-argument relations contained in the structures. The two rankings are interpolated into a *global* ranking to predict the structure preferred by

people. Difficulty is predicted with respect to the global ranking and the two local rankings, via two cost functions: *Conflict cost* quantifies the processing difficulty incurred in situations where the input yields conflicting evidence for which analysis to prefer, while *revision cost* accounts for the processing difficulty caused by abandoning a preferred interpretation of the input and replacing it with another.

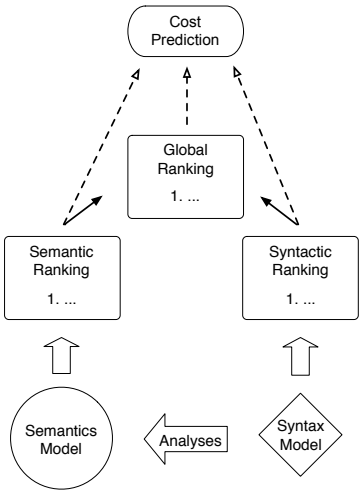


Figure 3. The architecture of the SynSem-Integration model, from Pado *et al.* (2009)

The integration of plausibility into a probabilistic sentence processing architecture, enables Pado *et al.* (2009) to model the findings of eight reading-time studies, covering four ambiguity phenomena, including the NP/S ambiguity (2), PP attachment (6), and reduce-relative clauses (7), discussed earlier. Crucially, each of the modeled studies revealed the on-line influence of plausibility on disambiguation during human parsing. While previous models have accounted for some of these findings with *hand-crafted* models for specific ambiguities (McRae *et al.*, 1998; Narayanan & Jurafsky, 1998; Tanenhaus *et al.*,

2000), the SynSem-Integration model offers a wide-coverage model, trained on syntactically and semantically annotated corpora, avoiding the need to specify the set of relevant constraints and their probabilities by hand for each new phenomenon to be modeled.

5 Connectionist Models of Sentence Processing

Connectionist networks, also called *artificial neural networks* (see Chapter 10), offer an alternative computational paradigm with which to model cognitive development and processing. While there is a tremendous variety of network architectures, most derive their inspiration from an abstraction of how the brain works: massively interconnected simple processing units (often called neurons) that operate in parallel. These units are usually grouped into *layers*, that themselves are an abstraction of the functional organization of the brain. Connectionist models of human sentence processing are attractive in that they inherit the experience-based behavior of probabilistic models, as a direct consequence of their ability to learn. Connectionist systems are typically trained through the adjustment of connection strengths in response to repeated exposure to relevant examples, thereby providing an integrated account of how both acquisition and subsequent processing are determined by the linguistic environment.

Connectionist models have been successfully applied to various aspects of human lexical processing, and crucially emphasize the importance of experience, specifically word frequency, for both learning and subsequent processing (Plunkett & Marchman, 1996; Christiansen & Chater, 1999a, 2001). Recent research, however, has also seen the emergence of sentence-level connectionist models which place similar emphasis on distributional information.

5.1 Simple Recurrent Networks

Simple recurrent networks (SRNs) provide an elegant architecture for learning distributional regularities that occur in sequential inputs (Elman, 1990). SRNs process patterns (vectors) rather than symbolic representations. SRNs process

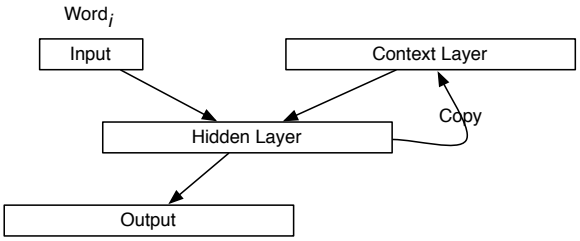


Figure 4. A simple recurrent network.

sentences one word at a time, with each new input word represented in the *input layer* and interpreted in the context of the sentence processed so far — represented by the *context layer*, which is simply a copy of the *hidden layer* from the previous time step (see Figure 4. The input layer and context layer are integrated and compressed into the hidden layer, enabling the network to incrementally develop a distributed representation of an unfolding sentence. Layers, in turn, may be partitioned into *assemblies* that are dedicated to specific functional tasks. The *output layer* contains patterns that the SRN has been trained to compute by providing targets for each output assembly. The target output may be some desired syntactic or semantic representation, but often SRNs are simply trained to predict the next word of the input, much like a probabilistic language model (Chapter 3). Each unit in the network receives a weighted sum of the input units feeding into it, and outputs a value according to an activation function that generally is nonlinear in order to bound the output value in an interval such as [0,1], such as the *logistic function*, $\sigma(x) = (1 + e^{-x})^{-1}$.

SRNs are trained by providing an input sequence and a set of targets into which the network should transform the input sequence. The standard training algorithm is *backpropagation*, an optimization technique that uses

error signals derived from the difference between the network's output and target to update the network weights to more closely approximate the targets on the next round of updates (Rumelhart *et al.*, 1986). The weights between units could themselves grow without bound during training, but an input vector $\mathbf{x}$ transformed by the matrix of weights $\mathbf{W}$ to produce an output vector $\mathbf{y}$ that has been passed through the activation function σ ensures $\mathbf{y}$ remains bounded. In sum, for each pair of layers connected by a weight matrix, the output vector can be calculated simply as $\mathbf{y} = \sigma(\mathbf{Wx})$.

One of the strengths of SRNs is that they can be trained on unannotated linguistic data, using the so-called *prediction task*: the network is presented with sentences, one word at a time, and is trained to output the *next* word in the sentence. To do this successfully, the network must learn those probabilistic and structural properties of the input language that constrain what the next word can be. The key insight of SRNs is the use of the context layer, which provides an exact copy of the hidden unit layer from the previous time step. This allows the network to combine information about its state at the previous time step with the current input word when predicting what words can follow. SRNs have been successfully trained on simplified, English-like languages based on grammars which enforce a range of linguistic constraints such as verb frame, agreement, and embedding (Elman, 1991). To learn these languages, the network must not only learn simple adjacencies, like the fact that "*the*" can be followed by "*boy*", but not "*ate*", but also long distance dependencies. Consider the following sentence initial fragment:

(8) "*The boy that the dog chased* ___."

In predicting what word will follow *chased*, the network has to learn that, although *chased* is a transitive verb, it cannot be followed by a noun phrase in this context, because of the relative clause construction. Rather it must be followed by the main verb of the sentence, which must further be singular since *the boy* is a singular subject. Interestingly, however, SRNs do exhibit limitations which appear to correspond well with those exhibited by people, namely in the processing of center-embedding constructions, as discussed in Section 3 (Christiansen & Chater, 1999b; MacDonald & Christiansen, 2002).

As noted, SRNs provide a model of incremental sentence processing, in which the network is presented with a sentence, word-by-word, and at each point attempts to predict which words will follow. Not only are SRNs able to learn complex distributional constraints with considerable success, they do so in a manner which reflects the relative frequencies of the training corpus. When the SRN is presented with the initial words of some sentence, $w_1 \ldots w_i$, it activates outputs corresponding exactly to those words which could come next. Furthermore, the *degree* of activation of the next word w_{i+1} corresponds closely to the conditional probability, as would be computed by a statistical language model as shown in Equation 8 (see Section 4 above, and Chapter 3).

$$P(w_{i+1}|w_1 \ldots w_i) = \frac{f(w_1 \ldots w_{i+1})}{f(w_1 \ldots w_i)} \tag{8}$$

Here, $f(w_1 \ldots w_{i+1})$ and $f(w_1 \ldots w_i)$ are the training corpus frequencies for the word sequences $w_1 \ldots w_{i+1}$ and $w_1 \ldots w_i$, respectively. The SRN thus not only predicts which words can follow, but it also the likelihood of each of those words, based on the conditional probabilities of those words in the training corpus.

One fundamental criticism of SRNs, however, is that there is only indirect evidence that syntactic structure is truly being acquired, at least in the conventional sense. Indeed, it has been argued that, although the language used to train the SRN was generated by a context-free grammar, the network may only be learning a weaker, probabilistic finite-state approximation in (8), rather than the true hierarchical structure of the language (Steedman, 1999). The lack of any explicit symbolic syntactic representation in SRNs also makes it difficult to model empirical evidence concerning the processing of syntactic ambiguity, since such ambiguity is predicated on the notion that two or more distinct hypotheses about the structure of the sentence must be distinguished during processing. The Visitation Set Gravitation model of Tabor *et al.* (1997), however, shows how reading times can be derived from a post-hoc analysis of a trained SRN. This analysis yields a landscape of *attractors* — points in multi-dimensional space that are derived from the hidden-unit activations, and which correspond to particular sentence structures. By observing how long it takes a particular hidden unit state (representing a word along with its left-context) to *gravitate* into an attractor (possibly representing a kind of semantic integration), Tabor et al. obtain a measure of the work a comprehender does integrating a word into a developing analysis.

Recursive neural networks (RNNs) Costa *et al.* (2003) can be seen as addressing Steedman (1999)'s criticism by developing an explicit model of structure disambiguation processes. RNNs are trained on a complete hierarchical representation of a syntactic tree, which is encoded in a multi-layer feed forward network in which the inputs represent the daughters and the output is the mother of a branch in the tree. The network is trained by exposing it recursively, from the leave of the tree, to each branch of the tree until the root

node is reached. The encoding of the root node thus represents an encoding of the entire tree. This enables training of the network using a parsed corpus (the Penn Treebank (Marcus *et al.*, 1993)), in which the network learns to make incremental parsing decisions, as in the relative clause attachment ambiguity shown in Fig. 4.2. Just as the SRN estimates the conditional probability of the next word given the words seen so far, the RNN estimates the conditional probability of each possible attachment for the current word, given the tree that has been built up to that point. The model therefore resembles a probabilistic parser, with the exception that RNNs are crucially able to learn *global* structural preferences (Sturt *et al.*, 2003), which standard PCFG models are not. RNNs can be seen as an implementation of the Mitchell *et al.* (1995)'s *Tuning Hypothesis* (Section 4.2), in that they are trained solely on syntactic structure, and not specific lexical items. One clear limitation of this approach, however, is that it does not account for lexical preferences or other kinds of non-structural biases (but see Costa *et al.* (2005) for discussion of some enhancements to this approach).

One recent SRN-based model has also sought to model aspects of visually-situated language understanding, as revealed by the *visual worlds* experiments (see end of Section 2.2). Mayberry *et al.* (2009) build on the theoretical proposal of Knoeferle & Crocker (2007), claiming that utterance-mediated attention in the visual context is not only driven by incremental and anticipatory linguistic processing, but crucially that it is this modulation of visual attention that underpins the rapid influence of the relevant visual context on comprehension — which they dub the *coordinated interplay account* (CIA). Mayberry *et al.* (2009)'s CIANet is based on a simple recurrent network (SRN; Elman, 1990) that produces a case-role interpretation of the input utterance. To al-

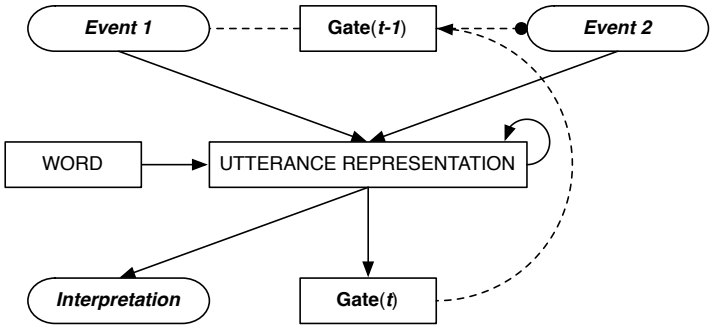


Figure 5. CIANet: A network featuring scene-language interaction with a basic attentional gating mechanism to select relevant events in a scene with respect to an unfolding utterance.

low visual input, CIANet incorporates an additional input representation of a scene as (optional) visual context for the input utterance. Scenes contain two events, only one of which is relevant to the input utterance, where each of the two scene events has three constituents (*agent, action* and *patient*) that are propagated to the SRN's hidden layer through shared weights (representing a common post-visual-processing pathway).

In line with the language-mediated visual attention mechanisms of the CIA, the unfolding linguistic input to CIANet modulates the activation of the relevant scene event based on the unfolding interpretation that is represented in the hidden layer. A gating vector implements the attentional mechanism in CIANet, and is multiplied element-wise with the corresponding units in each of the three lexical representations (agent, action, and patient) of one event (see Figure 5). Each unit of the gate is subtracted from 1.0 to derive a vector-complement that then modulates the second event. This means that more attention to one event in the model entails less attention to another. In this way, as the sentence is processed — possibly referring to the characters or actions in one of the scene events — the relevant is activating by the

gating vector, causing it to have a greater influence on the unfolding interpretation. The resulting network was shown to model the on-line influence of scene events on comprehension (Knoeferle *et al.*, 2005, Experiment 1), and the relative priority of depicted events versus stereotypical knowledge (Knoeferle & Crocker, 2006, Experiment 2), with the gating vector providing a qualitative model of experimentally observed visual attention behaviour. While the linguistic coverage of this model is currently limited to simple sentence structures, it is currently the only cognitive model of visually situated comprehension, and associated gaze behavior (but see Roy & Mukherjee (2005) for a psycholinguistically inspired account of how visual processing can influence speech understanding).

6 Hybrid Models

Within computational psycholinguistics, hybrid models can broadly be seen as identifying that class of architectures that combine explicit symbolic representations of linguistic structure and constraints with the use of connectionist inspired constraint-satisfaction and competitive activation techniques. Typically the goal of such approaches is to combine the transparent use of symbolic linguistic representations, which are absent in pure connectionist architectures, with the kinds of distributed, competitive and graded processing mechanisms that are absent in purely symbolic approaches. One early example is Stevenson (1994)'s CAPERS model, in which each word projects its phrasal structure as it is encountered, and initially all possible connections with the left-context are considered. Each possible attachment is assigned an activation, based on the extent to which is satisfies or violates lexical and syntactic constraints. Each node in the structure also has a limited amount of activation it can assign to it's connections, such that as some connections gain in strength, activation is taken away from others. The parser *iterates* until it stabilizes on a single, well-formed syntactic parse as each word is input. Vosse & Kempen (2000) propose a related model of parsing, based on a lexicalized grammar, in which possible *unification links* between words are graded and compete via lateral inhibition (see also Tabor & Hutchins (2004)'s SOPARSE for a related model, and more general discussion of this approach). The resulting model not only accounts for a range of standard parsing phenomena, but also the behavior found in some aphasic speakers.

As mentioned at the end of Section 3.1, constraint-based models of sentence disambiguation (MacDonald *et al.*, 1994; Tanenhaus *et al.*, 2000) deny that syntactic processes have any distinct modular status with the human lan-

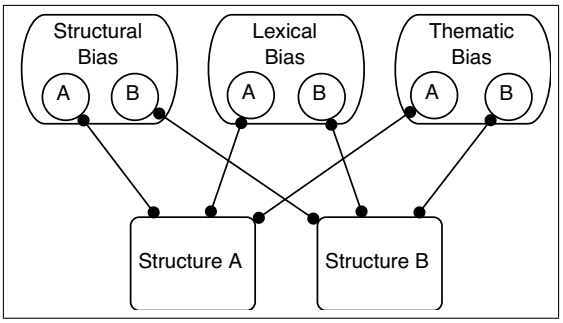


Figure 6. The competitive integration model (Spivey-Knowlton & Sedivy, 1995)

guage processor, rather assuming that all relevant constraints are integrated during processing. Such constraint-based accounts exploit the symbolic representations of linguistic constraints in combination with the use of competition-based constraint-satisfaction techniques (MacDonald *et al.*, 1994). The *competitive integration model* (Spivey-Knowlton & Sedivy, 1995; Spivey & Tanenhaus, 1998), for example, emphasizes the interaction of various heterogeneous linguistic constraints in resolving syntactic ambiguity, each with its own bias, (see Fig. 6) to be combined in deciding between several structural interpretations. For example, one might identify a general structural bias (as proposed by the Tuning Hypothesis, Section 4.2), a lexical verb frame bias, and perhaps a thematic-bias (e.g., the plausibility of either structure). McRae *et al.* (1998) proposed that the bias be established using experience-based measures: either corpus frequencies (e.g., for the structural and lexical constraint), or completions norms (e.g., for the thematic constraint). Once the relevant linguistic constraints are stipulated, the model allows two kinds of parameters to be set (Tanenhaus *et al.*, 2000): (i) the *weight* of each constraint, e.g., structural, lexical, and thematic, must be determined, and (ii) for each constraint, its *bias* towards Structure A versus Structure B must be established.

Once the parameters for the model have been determined, reading times are modeled by observing the time it takes the model to settle on a preferred structure as the different constraints compete. Informally, activation is propagated from each constraint to a particular structural candidate in accordance with the constraints bias. The activation of a given structure is then computed as the weighted sum of the activations from each constraint. This activation is then propagated back to the constraints, and another iteration is performed. Once the activation for a particular structure exceeds a specific threshold value, the system begins processing the next word. The number of iterations required for processing each word is then directly linked to the reading times observed during self-paced reading. McRae *et al.* (1998) demonstrate how the model can be used successfully to predict reading times for reduced-relative clauses, as a function of their semantic plausibility, as in example (7), above.

One short-coming of this approach, however, is that the model separates the mechanism which builds interpretations from the mechanism which chooses the interpretation. While independent modeling of the constraint reconciliation mechanisms might simply be viewed as *abstracting away* from the underlying structure building processes, the approach implies that structure building itself does not contribute to processing complexity, since it is the constraint integration mechanisms alone that determines reading times. Furthermore !!! ... with each disambiguation phenomena being modeled by a separate instance of the model (Tanenhaus *et al.*, 2000). Additionally, it might be argued that the number of degrees of freedom (both experience-based bias, and the weights of the constraints) reduces the predictive power of such models. Further empirical challenges to such constraint satisfaction models have

also been made (Frazier, 1995; Binder *et al.*, 2001) (but see also (Green & Mitchell, 2006)).

7 Concluding Remarks

The challenges of natural language understanding are daunting. Language is inherently complex — drawing on different levels of linguistic competence, as well as world and contextual knowledge — while also being highly ambiguous. That people are nonetheless able to comprehend language accurately and in real-time is a remarkable feat that is unmatched by any artificial system. Computational psycholinguistics is concerned with modeling how people achieve such performance, and seeks to develop implemented models of the architectures, mechanisms and representations involved. The approaches are diverse, ranging from purely symbolic accounts, to neurally inspired connectionist approaches, with hybrid and probabilistic models occupying the landscape in between. For reasons of space, we have focused our attention here on models of sentence processing, leaving aside models of lexical access (McClelland & Elman, 1986; Norris, 1999; Norris *et al.*, 2000). Equally, we have not addressed the topic of language acquisition, which is concerned with how our linguistic knowledge emerges as a consequence of linguistic experience. While the goals of acquisition and processing models differ with respect to the kinds of empirical data they attempt to explain, ultimately it is essential that models of adult sentence comprehension be the plausible end result of the acquisition process. The increasing dominance of experience-based models of language processing, whether connectionist and probabilistic, holds promise for a uniform and possibly even integrated account of language acquisition and adult performance (Chater & Manning, 2006). Indeed, language learning drove the early development of connectionist models of lexical and syntactic acquisition(Rumelhart & McClelland, 1987; Elman, 1990) which now figure prominently in computational psycholinguistics (Tabor *et al.*, 1997; Christiansen & Chater, 1999b,

2001; Mayberry *et al.*, 2009). Probabilistic, especially Bayesian, approaches have also been applied to problems of learning argument structure (Alishahi & Stevenson, 2008), syntax (Clark, 2001), and semantics (Niyogi, 2002). Not surprisingly, however, many models of acquisition emphasize the role of visual scene information(see also Siskind (1996)). Knoeferle & Crocker (2006) argue that this may explain the priority of visual context in adult sentence processing — as modeled by the CIANet architecture (Mayberry *et al.*, 2009) — further demonstrating the kind of synergy that may be possibly between acquisition and processing theories in future.

Virtually all modern accounts of sentence understanding share the assumption that language processing is highly incremental, with each encountered word being immediately integrated into an interpretation of what has been read or heard so far. Even this assumption, however, has been recently challenged by experimental findings suggesting that comprehension processes may build interpretations which make sense *locally*, even when they are ungrammatical with respect to the entire preceding-context (Tabor & Hutchins, 2004). Nonetheless, incrementality is almost certainly the rule, even if there are occasional exceptions. Indeed, there is an increasing emphasis on the role of *predictive* mechanisms in parsing, to explain the wealth of experimental findings that people not only processing language incrementally, but in fact actively generate hypothesis about the words they expect to follow. Much in the way that statistical language models assign probabilities to the words that may come next, both probabilistic (Hale, 2001, 2003; Levy, 2008) and connectionist (Elman, 1990, 1991; Mayberry *et al.*, 2009; Crocker *et al.*, in press) psycholinguistic models potentially offer natural explanations of predictive behavior in people.

There remain some issues which truly distinguish competing theories. For example whether or not people actively consider multiple interpretations in the face of ambiguity, or adopt a single one, backtracking to some alternative when necessary. Similarly, the degree of modularity is often viewed as a defining characteristic. While it has proven challenging to decide definitively among these positions empirically, there is increasing consensus that language comprehension mechanisms must support the rapid and adaptive integration of virtually all relevant information — linguistic and world knowledge, as well as discourse and visual context — as reflected by incremental and predictive comprehension behavior (see Crocker *et al.* (in press), for an overview of relevant empirical findings).

Finally, some models that appear quite different superficially may simply be offering accounts of processing at different levels of abstraction. Connectionist and probabilistic approaches most often share the idea that language understanding is an optimized process which yields, for example, the most likely interpretation for some input based on prior experience. Thus SRNs make very similar behavioral predictions to probabilistic language models based on n-grams or PCFGs. Typically, however, connectionist models intend to provide and account at the algorithmic or even implementation level in Marr's terms (recall Section 2.1), while probabilistic approaches may be construed as theories at the higher, *computational*, level. That is, while connectionist learning and distributed representations are postulated to have some degree of biological plausibility, the parsing and training mechanisms of probabilistic models typically are not. Hybrid architectures occupy a middle ground, combining explicitly stipulated symbolic representations with connectionist-inspired processing mechanisms.

The emergence of experience-based approaches represents a major milestone for computational psycholinguistics, resulting in models that offer broader coverage (Crocker & Brants, 2000) and rational behavior (Chater *et al.*, 1998; Crocker, 2005), while also explaining a wide-range of experimentally observed frequency effects (Jurafsky, 2002). As can be seen from the models discussed in this article, however, there is a tendency to isolate language processing from other cognitive processes such as perception and action. As such, computational models are lagging behind emerging theories of situated and embodied language processing, which emphasize the interplay and overlap of language, perception and action (Barsalou, 1999; Fischer & Zwaan, 2008; Spivey & Richardson, 2009). The CIANet model (Mayberry *et al.*, 2009) is a one attempt to model visually situated comprehension, thereby also connecting with situated language learning models, but computational psycholinguistics still lags behind current experimental results and theoretical claims concerning the integration of language with other cognitive systems. Future developments in this direction will likely connect with models of language acquisition, and ultimately contribute to a better understanding of the origins of the human capacity for language.

References

Abney, Steven Paul (1989), A computational model of human parsing, *Journal of Psycholinguistic Research* 18:129–144.

Abney, Steven Paul & Mark Johnson (1991), Memory requirements and local ambiguities of parsing strategies, *Journal of Psycholinguistic Research* 20:233–250.

Alishahi, Afra & Suzanne Stevenson (2008), A Computational Model of Early Argument Structure Acquisition, *Cognitive Science* 32(5):789–834.

Allopenna, Paul D., James S. Magnuson, & Michael K. Tanenhaus (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.

Altmann, Gerry T. M. & Yuki Kamide (1999), Incremental interpretation at verbs: restricting the domain of subsequent reference, *Cognition* 73:247–264.

Anderson, John R. (1990), *The adaptive character of thought.*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Anderson, John R., Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, & Yulin Qin (2004), An integrated theory of the mind, *Psychological Review* 111:1036–1060.

Barsalou, L. W. (1999), Perceptual and symbol systems, *Behavioural and Brain Sciences* 22:577–609.

Baum, Leonard E. (1972), An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, *Inequalities* 3(1):1–8.

Bever, Tom (1970), The cognitive basis for linguistic structures, in J. R. Hayes (ed.), *Cognition and the development of language*, Wiley, New York, (279–362).

Binder, Katherine S., Susan A. Duffy, & Keith Rayner (2001), The effects of thematic fit and discourse context on syntactic ambiguity resolution, *Journal of Memory and Language* 44:297–324.

Bornkessel, Ina & Matthias Schlesewsky (2006), The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages, *Psychological Review* 113:787–821.

Brants, Thorsten (1999), Cascaded markov models, in *9th Conference of the European Chapter of the Association for Computational Linguistics (EACL '99), June 8–12*, Bergen, (118–125).

Brants, Thorsten & Matthew W. Crocker (2000), Probabilistic parsing and psychological plausibility, in *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken/Luxembourg/Nancy, (111–117).

Brysbaert, Marc & Don C. Mitchell (1996), Modifier attachment in sentence parsing: Evidence from Dutch, *Quarterly Journal of Experimental Psychology* 49A(3):664–695.

Chater, Nick, Matthew W. Crocker, & Martin J. Pickering (1998), The rational analysis of inquiry: The case for parsing, in Nick Chater & M. Oaksford (eds.), *Rational Analysis of Cognition*, Oxford University Press, Oxford, (441–468).

Chater, Nick & Christopher D. Manning (2006), Probabilistic models of language processing and acquisition, *Trends in Cognitive Science* 10(7):335–344.

Christiansen, Morten H. & Nick Chater (1999a), Connectionist natural language processing: The state of the art, *Cognitive Science* 23:417–437.

Christiansen, Morten H. & Nick Chater (1999b), Toward a connectionist model of recursion in human linguistic performance, *Cognition Science* 23(2):157–205.

Christiansen, Morten H. & Nick Chater (2001), Connectionist psycholinguistics: Capturing the empirical data, *Trends in Cognitive Sciences* 5(2):82–88.

Clark, Alexander (2001), *Unsupervised Language Acquisition: Theory and Practice*, Ph.D. thesis, COGS, University of Sussex.

Corley, Steffan & Matthew Crocker (2000), The modular statistical hypothesis: Exploring lexical category ambiguity, in Matthew Crocker, Martin Pickering, & Charles Clifton (eds.), *Architectures and Mechanisms for Language Processing*, Cambridge University Press, Cambridge, (135–160).

Costa, Fabrizio, Paolo Frasconi, Vincenzo Lombardo, & Giovanni Soda (2003), Towards incremental parsing of natural language using recursive neural networks, *Applied Intelligence* 19:9–25.

Costa, Fabrizio, Paolo Frasconi, Vincenzo Lombardo, & Giovanni Soda (2005), Ambiguity resolution analysis in incremental parsing of natural language, *IEEE Transactions on Neural Networks* 16:959–971.

Crain, Stephen & Mark Steedman (1985), On not being led up the garden path: the use of context by the psychological parser, in D. Dowty, L. Karttunnen, & A. Zwicky (eds.), *Natural language parsing*, Cambridge University Press, Cambridge, MA, (320–358).

Crocker, Matthew & Steffan Corley (2002), Modular architectures and statistical mechanisms: The case from lexical category disambiguation, in Suzanne Stevenson & Paola Merlo (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*, John Bejamins, Amsterdam, (157–180).

Crocker, Matthew W. (1996), *Computational Psycholinguistics: An Interdisciplinary Approach to the Study of Language*, Kluwer, Dordrecht.

Crocker, Matthew W. (1999), Mechanisms for sentence processing, in Simon Garrod & Martin J. Pickering (eds.), *Language Processing*, Psychology Press, London, (191–232).

Crocker, Matthew W. (2005), Rational models of comprehension: Addressing the performance paradox, in Anne Cutler (ed.), *Twenty-First Century Psycholinguistics*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, (363–380).

Crocker, Matthew W. & Thorsten Brants (2000), Wide-coverage probabilistic sentence processing, *Journal of Psycholinguistic Research* 29(6):647–669.

Crocker, Matthew W. & Frank Keller (2006), Probabilistic grammars as models of gradience in language processing, in G. Fanselow, C. Fry, R. Vogel, & M. Schlesewsky (eds.), *Gradience in Grammar: generative perspectives*, Oxford University Press, Oxford, UK; New York, (227–245).

Crocker, Matthew W., Pia Knoeferle, & Marshall Mayberry (in press), Situated sentence comprehension: The coordinated interplay account and a neurobehavioral model, *Brain and Language* .

Cuetos, Fernando & Don C. Mitchell (1988), Cross-linguistic differences in parsing: Restrictions on the late closure strategy in Spanish, *Cognition* 30:73–105.

Cuetos, Fernando, Don C. Mitchell, & Martin M. B. Corley (1996), Parsing in different languages, in Manuel Carreiras, J. García-Albea, & N. Sabastián-Gallés (eds.), *Language Processing in Spanish*, Lawrence Erlbaum Associates, Mahwah, NJ, (145–189).

Duffy, Susan A., Robin K. Morris, & Keith Rayner (1988), Lexical ambiguity and fixation times in reading, *Journal of Memory and Language* 27:429–446.

Elman, Jeffrey L. (1990), Finding structure in time, *Cognition Science* 14(2):179–211.

Elman, Jeffrey L. (1991), Distributed reprentations, simple recurrent networks and grammatical structure, *Machine Learning* 9:195–225.

Federmeier, Kara (2007), Thinking ahead: The role and roots of prediction in language comprehension, *Psychophysiology* 44:491–505.

Fillmore, C. J. (1968), The case for case, in Emmon Bach & Robert T. Harms (eds.), *Universals in Linguistic Theory*, Holt, Reinhart and Winston, New York, (1–88).

Fillmore, Charles, Christopher Johnson, & Miriam Petruck (2003), Background to framenet, *International Journal of Lexicography* 16:235–250.

Fischer, Martin H. & Rolf A. Zwaan (2008), Embodied language: A review of the role of the motor system in language comprehension, *Quarterly Journal of Experimental Psychology* 61(6):825–850.

Fodor, Jerry (1983), *Modularity of mind*, MIT Press, Cambridge, MA.

Ford, Marilyn, Joan Bresnan, & Ronald Kaplan (1977), A competence-based theory of syntactic closure, in Joan Bresnan (ed.), *The mental representation of grammatical relations*, MIT Press, Cambridge. MA, (727–796).

Frazier, Lyn (1979), *On comprehending sentences: Syntactic parsing strategies*, Unpublished doctoral dissertation, University of Connecticut.

Frazier, Lyn (1995), Constraint satisfaction as a theory of sentence processing, *Journal of Psycholinguistic Research* 24:437–468.

Frazier, Lyn & Keith Rayner (1987), Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences, *Journal of Memory and Language* 26:505–526.

Friederici, Angela D. (2002), Towards a neural basis of auditory sentence processing, *Trends in Cognitive Science* 6(2):78–84.

Garnsey, Susan M., Neal J. Pearlmutter, Elisabeth M. Myers, & Melanie A. Lotocky (1997), The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences, *Journal of Memory and Language* 37(1):58–93.

Gibson, Edward (1991), *A computational theory of human linguistic processing: Memory limitations and processing breakdown*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

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

Gibson, Edward (2003), Linguistic complexity in sentence comprehension, in Philipp Strazny (ed.), *The Encyclopedia of Cognitive Science*, MacMillan, New York, (240–241).

Gibson, Edward & Neal J. Pearlmutter (1998), Constraints on sentence comprehension, *Trends in Cognitive Sciences* 2(7):262–268.

Green, Matthew J. & Don C. Mitchell (2006), Absence of real evidence against competition during syntactic ambiguity resolution, *Journal of Memory and Language* 55:1–17.

Grosjean, Francois (1980), Spoken word recognition processes and the gating paradigm, *Perception and Psychophysics* 28:267–283.

Hale, John (2001), A probabilistic earley parser as a psycholinguistic model, in *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA, (159??166).

Hale, John (2003), The information conveyed by words in sentences, *Journal of Psycholinguistic Research* 32(1):101–122.

Johnson-Laird, Philip N. (1983), *Mental Models*, Harvard University Press.

Juliano, Cornell & Michael K. Tanenhaus (1993), Contingent frequency effects in syntactic ambiguity resolution, in *The 15th Annual Conference of the Cognitive Science Society*, (593–603), 681–686.

Jurafsky, Dan (2002), Probabilistic modeling in psycholinguistics: Linguistic comprehension and production, in Rens Bod, Jennifer Hay, & Stefanie Jannedy (eds.), *Probabilistic Linguistics*, MIT Press, (39–96).

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

Jurafsky, Daniel (2003), Probabilistic modeling in psycholinguistics: Linguistic comprehension and production, in Rens Bod, Jennifer Hay, & Stefanie Jannedy (eds.), *Probabilistic Linguistics*, MIT Press, Cambridge, MA, (39–95).

Kimball, J. (1973), Seven principles of surface structure parsing in natural languages, *Cognition* 2:15–47.

Knoeferle, P. & M. W. Crocker (2006), The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking, *Cognitive Science* 30:481–529.

Knoeferle, P. & M. W. Crocker (2007), The influence of recent scene events on spoken comprehension: Evidence from eye movements, *Journal of Memory and Language* 57:519–543.

Knoeferle, Pia, Matthew W. Crocker, Christoph Scheepers, & Martin J. Pickering (2005), The influence of the immediate visual context on incremental thematic role-assignment, *Cognition* 95:95–127.

Kutas, M. & S. A. Hillyard (1980), Reading senseless sentences: brain potentials reflect semantic incongruity, *Science* 207:203–205.

Kutas, Marta & Steven A. Hillyard (1983), Event-related brain potentials to grammatical errors and semantic anomalies, *Memory and Cognition* 11:539–550.

Levy, Roger (2008), Expectation-based syntactic comprehension, *Cognition* 106(3):1126–1177.

Lewis, Richard L., Shravan Vasishth, & Julie A. Van Dyke (2006), Computational principles of working memory in sentence comprehension, *Trends in Cognitive Science* 10:447–454.

MacDonald, Maryellen C. (1993), The interaction of lexical and syntactic ambiguity, *Journal of Memory and Language* 32:692–715.

Macdonald, Maryellen C. (1994), Probabilistic constraints and syntactic ambiguity resolution, *Language and Cognitive Processes* 9:157–201.

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

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

Manning, Christopher D. & Hinrich Schütze (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

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

Marcus, Mitchell P., Beatrice Santorini, & Mary Ann Marcinkiewicz (1993), Building a large annotated corpus of English: The Penn Treebank, *Computational Linguistics* 19(2):313–330.

Marr, David (1982), *Vision*, W. H. Freeman and Company, San Francisco, CA, USA.

Matzke, Mike, Heinke Mai, Wido Nager, Jascha Rsseler, & Thomas Mnte (2002), The costs of freedom: an erp-study of non-canonical sentences, *Clinical Neuropsychology* 113:844–852.

Mayberry, Marty, Matthew W. Crocker, & P. Knoeferle (2009), Learning to attend: A connectionist model of the coordinated interplay of utterance, visual context, and world knowledge., *Cognitive Science* 33:449–496.

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

McRae, Ken, Michael J. Spivey-Knowlton, & Michael K. Tanenhaus (1998), Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension, *Journal of Memory and Language* 38:283–312.

Miller, George A. & Stephen Isard (1964), Free recall of self-embedded english sentences, *Information and Control* 7:292–303.

Mitchell, Don C., Fernando Cuetos, Martin Corley, & MArc Brysbaert (1995), Exposure-based models of human parsing: evidence for the use of coarse-grained (nonlexical) statistical records, *Journal of Psycholinguistic Research* 24:469–488.

Narayanan, Srini & Daniel Jurafsky (1998), Bayesian models of human sentencing processing, in *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, (752–757).

Niyogi, Sourabh (2002), Bayesian learning at the syntax-semantics interface, in *Proceedings of the 24th annual conference of the Cognitive Science Society*, (697–702).

Norris, Dennis (1999), Computational psycholinguistics, in Wilson & Keil (eds.), *MIT Encyclopedia of the Cognitive Sciences,*, MIT Press, Cambridge, MA, (168–169).

Norris, Dennis, James McQueen, & Anne Cutler (2000), Merging information in speech processing: Feedback is never necessary, *Behavioral and Brain Sciences* 23(3):299–370.

Osterhout, Lee & P. J. Holcomb (1992), Event-related brain potentials elicited by syntactic anomaly, *Journal of Memory and Language* 31:785–806.

Osterhout, Lee & Phillip J. Holcomb (1993), Event-related potentials and syntactic anomaly: evidence of anomaly brain potentials, *Psychophysioloy* 30:170–182.

Pado, Ulrike, Matthew W. Crocker, & Frank Keller (2009), A probabilistic model of semantic plausibility in sentence processing, *Cognitive Science* 33:794–838.

Pereira, Fernando (1985), A new characterization of attachment preferences, in *Natural Language Parsing—Psychological, Computational and Theoretical perspectives*, Cambridge University Press, (307–319).

van Petten, Cyma & Marta Kutas (1990), Interactions between sentence context and word frequency in event-related brain potentials, *Memory and Cognition* 18:380–393.

Pickering, Martin J., Charles Clifton, & Matthew W. Crocker (2000a), Architectures and mechanisms in sentence comprehension, in Matthew W. Crocker, Martin J. Pickering, & Charles Clifton (eds.), *Architectures and mechanisms for language processing*, Cambridge University Press, Cambridge, (1–28).

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

Plunkett, Kim & Virginia A. Marchman (1996), Learning from a connectionist model of the acquisition of the English past tense, *Cognition* 61(3):299–308.

Pritchett, B. L. (1992), *Grammatical competence and parsing performance*, University of Chicago Press, Chicago.

Rayner, Keith (1998), Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin* 124:372–422.

Rayner, Keith, Marcia Carlson, & Lyn Frazier (1983), The interaction of syntax and semantics during sentence processing: eye movements in the analysis of semantically biased sentences, *Journal of Verbal Learning and Verbal Behavior* 22:358–374.

Resnik, Philip (1992), Left-corner parsing and psychological plausibility, in *Proceedings of the Fourteenth International Conference on Computational Linguistics*, (191–197).

Roark, Brian (2001), *Robust Probabilistic Predictive Syntactic Processing: Motivations, Models, and Applications*, Ph.D. thesis, Brown University.

Roy, Deb & Niloy Mukherjee (2005), Towards situated speech understanding: visual context priming of language models, *Computer Speech and Language* 19:227–248.

Rumelhart, David E., Geoffrey E. Hinton, & Ronald J. Williams (1986), Learning internal representations by error propagation, in David E. Rumelhart & James L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, MIT Press, Cambridge, MA, (318–362).

Rumelhart, David E. & Jay L. McClelland (1987), Learning the past tenses of English verbs: Implicit rules or parallel distributed processing, *Mechanisms of language acquisition* :195–248.

Shannon, Claude E. (1948), A mathematical theory of communication, *Bell System Technical Journal* 27:379–423, 623–656.

Shapiro, Stuart C. (1992), Artificial intelligence, in Stuart C. Shapiro (ed.), *The Encyclopedia of Artificial Intelligence*, John Wiley & Sons, New York, (54–57).

Siskind, Jeffrey M. (1996), A computational study of cross-situational techniques for learning word-to-meaning mappings, *Cognition* 61(1-2):39–91.

Spivey, Michael J. & Daniel C. Richardson (2009), Language embedded in the environment, in P. Robbins & M. Aydede (eds.), *The Cambridge Handbook of Situated Cognition*, Cambridge University Press, Cambridge, UK, (382–400).

Spivey, Michael J. & Michael K. Tanenhaus (1998), Syntactic ambiguity resolution in discourse: modeling the effects of referential context and lexical frequency, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24:1521–1543.

Spivey-Knowlton, Michael & Julie Sedivy (1995), Resolving attachment ambiguities with multiple constraints, *Cognition* 55:227–267.

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

Stevenson, Suzanne (1994), Competition and recency in a hybrid network model of syntactic disambiguation, *Journal of Psycholinguistic Research* 23(4):295–322.

Stolcke, Andreas (1995), An efficient probabilistic context-free parsing algorithm that computes prefix probabilities, *Computational Linguistics* 21(2):165–201.

Sturt, Patrick, Fabrizio Costa, Vincenzo Lombardo, & Paolo Frasconi (2003), Learning first-pass structural attachment preferences using dynamic grammars and recursive neural networks, *Cognition* 88:133–169.

Sturt, Patrick, Martin J. Pickering, & Matthew W. Crocker (1999), Structural change and reanalysis diffulty in language comprehension: Is reanalysis the last resort?, *Journal of Memory and Language* 40(1):136–150.

Tabor, Whitney & Sean Hutchins (2004), Evidence for seof-organized sentence processing: Digging-in effects, *Journal of Experimental Psychology: Learning, Memory and Cognition* 30:431–450.

Tabor, Whitney, Cornell Juliano, & Michael K. Tanenhaus (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.

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard, & Julie C. Sedivy (1995), Integration of visual and linguistic information in spoken language comprehension, *Science* 268:1632–1634.

Tanenhaus, Michael K., John C. Trueswell, & J. E. Hanna (2000), Modeling thematic and discourse context effects with a multiple constraints approach: implications for the architecture of the language comprehension system, in M. W. Crocker, Martin J. Pickering, & C. Clifton (eds.), *Architectures and mechanism for language processing*, Cambridge University Press, Cambridge, (90–118).

Trueswell, J. C., Michael K. Tanenhaus, & C. Kello (1993), Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths, *Journal of Experimental Psychology: Learning, Memory and Cognition* 19:528–553.

Trueswell, John C. (1996), The role of lexical frequency in syntactic ambiguity resolution, *Journal of Memory and Language* 35:566–585.

Trueswell, John C. & Michael K. Tanenhaus (1994), Toward a lexicalist framework for constraint-based syntactic ambiguity resolution, in Charles Clifton, Lyn Frazier, & Keith Rayner (eds.), *Perspectives in Sentence Processing*, Lawrence Erlbaum, Hillsdale, NJ, (155–179).

Viterbi, Andrew J. (1967), Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Transactions on Information Processing* 13:260–269.

Vosse, Theo & Gerard Kempen (2000), Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar, *Cognition* 75:105–143.

Winograd, Terry (1983), *Language as a Cognitive Process*, Addison Wesley.