In press, Cognitive Science

A Probabilistic Model of Lexical and Syntactic Access and Disambiguation

Daniel Jurafsky
International Computer Science Institute &
Department of Linguistics
University of California at Berkeley
jurafsky@icsi.berkeley.edu

November 6, 1995

Abstract

The problems of *access* – retrieving linguistic structure from some mental grammar – and *disambiguation* – choosing among these structures to correctly parse ambiguous linguistic input – are fundamental to language understanding. The literature abounds with psychological results on lexical access, the access of idioms, syntactic rule access, parsing preferences, syntactic disambiguation, and the processing of garden-path sentences. Unfortunately, it has been difficult to combine models which account for these results to build a general, uniform model of access and disambiguation at the lexical, idiomatic, and syntactic levels. For example psycholinguistic theories of lexical access and idiom access and parsing theories of syntactic rule access have almost no commonality in methodology or coverage of psycholinguistic data. This paper presents a single probabilistic algorithm which models both the access and disambiguation of linguistic knowledge. The algorithm is based on a parallel parser which ranks constructions for access, and interpretations for disambiguation, by their conditional probability. Low-ranked constructions and interpretations are pruned through beam-search; this pruning accounts, among other things, for the garden-path effect. I show that this motivated probabilistic treatment accounts for a wide variety of psycholinguistic results, arguing for a more uniform representation of linguistic knowledge and for the use of probabilistically-enriched grammars and interpreters as models of human knowledge of and processing of language.

1 Introduction

The problems of *access* – retrieving linguistic structure from some mental grammar – and *disambiguation* – choosing among combinations of these structures to correctly parse ambiguous linguistic input – are fundamental to language understanding. Recently a number of computational cognitive models of these tasks have appeared, including models of lexical access, (Cottrell 1985; McClelland and Elman 1986), lexical disambiguation (Small and Rieger 1982), syntactic disambiguation (Abney 1989;

McRoy and Hirst 1990; Shieber 1983), the access and disambiguation of idioms (Wilensky and Arens 1980; van der Linden and Kraaij 1990; van der Linden 1992), and the difficulties in disambiguating garden-path sentences (Gibson 1991; Pritchett 1988; Henderson 1994).

It has proven difficult to combine these models to build a general, uniform model of access and disambiguation at the lexical, idiomatic, and syntactic levels. For example psycholinguistic theories of lexical access and idiom access and parsing theories of syntactic rule access have almost no commonality in methodology or coverage of psycholinguistic data. In part this has been due to an assumption that these levels, or at least the lexical and syntactic levels, are functionally distinct in the human language processor. In part a uniform model has been difficult to reconcile with traditional linguistic models, which have tended to split into phonological, syntactic, and semantic modules. And finally, uniformity has met with problems from the psycholinguistic data which have suggested quite different models for lexical and syntactic processing. For example, there is a great deal of evidence for *parallelism* in lexical processing (Swinney (1979), Tanenhaus et al. (1979), and Tyler and Marslen-Wilson (1982)). Thus in most models of lexical access, multiple candidate words are activated from the mental lexicon based on orthographic or phonological input. Psycholinguistic models of syntactic processing, however, have generally been serial, rather than parallel. That is, only one parse of the input is maintained at all times. One reason for this has been the need to account for the unacceptability of *garden-path* sentences like (1).

(1) # The horse raced past the barn fell. *(Bever 1970)*

As we will see in §4, sentences like (1) are difficult because they are locally ambiguous at some point in the processing (for example between a main-verb and reduced-relative reading of the word 'raced'). In most syntactic accounts of this data, the serial nature of the parser requires only one of these interpretations to be chosen, and local heuristics cause the incorrect main verb reading to be chosen. These serial models of syntactic processing are incompatible with the parallel models of lexical processing.

Finally, while robust *frequency* effects have long been noted in lexical processing, neither linguistic theories of syntax, nor the parsing algorithms which use them to build parses, have suggested methods for dealing with probabilistic effects in syntax.

Despite these problems and differences, most researchers agree that at least some features of the language processing architecture are uniform across levels. For example there is evidence for *on-line* processing at the lexical, idiomatic, and syntactic levels, including evidence from comprehension (Marslen-Wilson 1975; Potter and Faulconer 1979), lexical disambiguation (Swinney 1979; Tanenhaus et al. 1979; Tyler and Marslen-Wilson 1982; Marslen-Wilson et al. 1988), pronominal anaphora resolution (Garrod and Sanford 1991; Swinney and Osterhout 1990), verbal control (Boland et al. 1990; Tanenhaus et al. 1989), and gap filling (Crain and Fodor 1985; Stowe 1986; Carlson and Tanenhaus 1987; Garnsey et al. 1989; Kurtzman et al. 1991).

But more fundamentally, recent years have seen a convergence among linguists and psychologists toward less modular, more interactive models of language and language processing. The rise of unification-based linguistic theories (Sag et al. 1985; Bresnan

1982; Pollard and Sag 1987) has led to the ability to represent information at every level of representation with the same formalism. Tanenhaus and associates (Spivey-Knowlton et al. 1993; Trueswell and Tanenhaus 1991; Tanenhaus et al. 1989) have argued that semantic aspects of lexical information such as verbal semantic valence play an early role in syntactic parsing, and that semantic information plays a role in garden-path effects. Finally, MacDonald (1993) and Kawamoto (1993) have noted a number of similarities between lexical and syntactic disambiguation, and suggested that a uniform constraint-based mechanism might be built to account for disambiguation across linguistic levels.

In this paper we follow these recent directions to propose a single computational architecture that models both access and disambiguation, accounting for psycholinguistic results on lexical and idiomatic access, lexical and syntactic disambiguation, and garden-path sentences. The model is based on a preference for *coherence* in interpretation. Intuitively, constructions are more likely to be accessed when they are more coherent with the previous context, and interpretations are preferred over other interpretations when they are more coherent. Significantly, we show that this intuition of coherence can be formalized with a probabilistic foundation. While linguistic grammars which include frequency information date at least back to Ulvestad's (1960) work on subcategorization, only recently has a body of techniques become common for dealing with stochastic grammars (Baker 1979; Jelinek and Lafferty 1991; Fujisaki et al. 1991) in which every rule of the grammar is augmented with a conditional probability. Recent studies have applied these probabilities to linguistic theories like LTAG (Resnik 1992), and have shown the use of such probabilistic grammars for modeling synchronic change (Tabor 1993) and learning (Stolcke 1994; Stolcke and Omohundro 1993). In this paper we show that a probabilistic model differs from the frequency-based models traditional in psycholinguistics and argue that true probabilities are essential for a cognitive model of sentence processing.

The algorithm is designed in the framework of *Construction Grammar* (Fillmore et al. 1988; Kay 1990; Lakoff 1987; Goldberg 1991; Goldberg 1992; Koenig 1993; Lakoff 1993; Lambrecht 1995), a unification-based theory in which the mental lexicon, idiom list, and grammar are represented as a uniform collection of *grammatical constructions*. Construction grammar is a sign-based theory, like HPSG (Pollard and Sag 1987), in which each construction represents well-formedness conditions across various domains of linguistic knowledge, including phonological, morphological, syntactic, semantic, and pragmatic domains. Although we describe the algorithm in the construction grammar framework, it can be applied to any sign-based unificational linguistic theory which allows the expression of predicate valence (HPSG, cognitive grammar, Montague Grammar, categorial grammar, RRG, probably LFG) simply by the addition of the appropriate probabilities.

The model consists of a parallel parser augmented with probabilities and a pruning heuristic which models both access and disambiguation. The model addresses access by unifying lexical access, syntactic rule access, and idiom access. In this **Bayesian access** or **evidential access** algorithm, phonological, syntactic and semantic evidence, top-down as well as bottom-up, is integrated to determine which constructions to access, by computing the conditional probability of the construction given each piece of

evidence. Lexical, idiomatic, or syntactic constructions are activated in parallel when the weight of evidence passes a threshold. Our model of construction disambiguation, the **local coherence** model, similarly conflates lexical disambiguation, idiom disambiguation, syntactic disambiguation, and semantic disambiguation. The algorithm again computes the conditional probability of each interpretation given the input, and ranks interpretations according to these probabilities. The algorithm prunes these interpretations to keep the search space manageable; a by-product of this pruning is the garden-path effect, which we model as the pruning of what turns out to be the correct parse.

The actual probability computation involves two sources of expectations, *constituent* and *valence* information. Constituent expectations arise from the phrase-structural skeleton of the grammar. Each context-free rule is annotated with a conditional probability that the left side will expand to the right side. These probabilities place a distribution over the kind of daughters each phrasal node requires. Valence, or subcategorization expectations are associated with verbs and other predicates. Each predicate may express constraints on the syntactic, thematic, or semantic form of their arguments. For each such argument, the predicate is also annotated with a conditional probability that the argument will be filled.

We draw three conclusions from this work. First, we show a unified model of a number of psycholinguistic phenomena, including parse preferences, explaining the difficulty of garden-path sentences, and results on lexical and idiom access. The model requires only simple probabilistic augmentations to concepts of phrase structure and valence that already exist in most modern grammars and theories of sentence processing. Second, we argue that these results demonstrate the need for probabilistic augmentation to grammatical theories if they are to be able to deal with a broad range of phenomena as a processing model. In this vein our model is part of a larger effort involving computational implementations of linguistic theory (Jurafsky 1992a) and large-scale implementations of probabilistic augmentations to linguistic models for speech recognition (Jurafsky et al. 1995b; Jurafsky et al. 1995a; Tajchman et al. 1995).

Finally, the idea that lexical, idiomatic, syntactic, and semantic structures are uniformly represented and processed contrasts sharply with the traditional modular view, which holds that each of these kinds of linguistic knowledge is separately represented and processed. A number of linguists and psychologists have recently argued that an integrated view of linguistic structure and processing is necessary to account for linguistic data (Lakoff 1987; Langacker 1987; Taraban and McClelland 1988; Fillmore et al. 1988; McClelland et al. 1989; Goldberg 1991; Trueswell and Tanenhaus 1991; MacDonald 1993; Spivey-Knowlton et al. 1993; Spivey-Knowlton and Sedivy 1995; Lambrecht 1995). We hope that by showing that a single parallel probabilistic algorithm deals uniformly with lexical access data, idiom processing, parsing preferences, and garden-path data, to provide additional evidence that a uniform, non-modular theory of language representation and process is possible.

2 Architecture

Access and disambiguation are only a small part of a complete model of parsing. In order to maintain the advantages of computational efficiency in the face of complex recursive structure, we assume our parser is built with the standard dynamic programming (chart) parsing architecture. Then our access and disambiguation algorithms are implemented as a set of pruning heuristics that augment this standard algorithm. In the next sections we give an overview of the motivation for using pruning heuristics as a cognitive model; detailed arguments will be presented in §3 and §4. We then turn to our model of linguistic representation.

2.1 Parallel Parsing

As we will discuss in §3, psycholinguistic results on lexical access and idiom access, as well as recent results on syntactic processing and our own arguments in §4 on garden-path processing lead us to believe that the underlying architecture of the human language interpretation mechanism is parallel (although cf Frazier (1987)).

However many researchers have noted that without some special attempts at efficiency, the problem of computing parallel syntactic structures for a sentence can be quite complex. Church and Patil (1982) showed, for example, that the number of ambiguous phrase-structure trees for a sentence with multiple preposition-phrases was proportional to the Catalan numbers, while Barton et al. (1987) showed that the need for keeping long-distance agreement information and the need to represent lexical ambiguity together make the parsing problem for a grammar that represents such information NP-complete. If we expect our parsing algorithms to extend to spoken-language input, the ambiguity problem explodes even further, first because just allowing the input to be strings of phones rather than words adds phonetic parse and lexical segmentation ambiguity, and second because phonetic estimation itself is likely to require probabilistic, non-discrete inputs to the parsing algorithm.

All solutions to the problem of maintaining multiple parses of an ambiguous sentence, while still parsing in polynomial time, involve dynamic programming techniques. These methods, from the *well-formed substring table* (WFST) of Kuno (1965), to the *chart parsing* algorithm of Kay (1973) and the *Earley* algorithm of Earley (1970), all essentially trade memory for processing time. The parser stores the common subparts of multiple parses, allowing sub-parses to be represented only once, instead of once per parse tree.

We assume that the human parser will need to use some such dynamic programming algorithm for parsing control and for rule-integration.

2.2 Why Pruning?

However, the efficiency gained by dynamic programming algorithms for pure syntax may not generalize to the problem of interpretation. For example, if two parse trees both include an NP, the dynamic programming algorithm can simply store the NP once, because the internal structure of the NP is irrelevant to the global parse. But if two interpretations share the same NP, it may not be possible to store the NP only

once, because its internal structure, and particularly its semantic structure, is relevant to the interpretation, and may be needed by the interpreter to produce part of an on-line interpretation. Building the semantics of the NP into the interpretation may involve binding variables differently in the context of different interpretations. Although some semantic structure can most likely be shared, the sharing will not be as efficient as for syntactic structure.

However, there is another simple way to improve the computational tractability of parsing. The results of Church and Patil (1982) and Barton et al. (1987) rely on the fact that syntactic ambiguities in these parsers are not resolved until *after* the entire sentence has been parsed. It is the need to represent ambiguities for indefinite lengths of time in parsing that causes complexity. One way for a cognitive model of parsing to be computationally tractable, then, is to do some sort of on-line *pruning*. That is, rather than searching the entire space of possible interpretations for an input utterance, we rank our hypotheses and abandon the disfavored ones, or indeed find some way to avoid even suggesting (accessing) them.

Our models of access and disambiguation are based on this idea. As we will see in §3, in a standard model of parsing like bottom-up parsing, every possible rule which could possibly account for the input data is accessed. As the rule-base grows, this becomes quite a large set. If we would like our access algorithms to extend to speech processing, the non-determinism of phonetic input makes the number of possible structures which could be input even larger. In order to make the access problem tractable, we argue that it will need to do some sort of pruning, either by accessing many structures and then dropping some, or by accessing fewer structures in the first place.

While access pruning happens as linguistic structures are suggested from the mental grammar, disambiguation happens after structures are integrated together. Disambiguation is a kind of pruning by definition, selecting an interpretation which is better than other interpretations on some metric. In order to avoid the computational complexity of maintaining ambiguities over long distances, our model disambiguates on-line by pruning unlikely interpretations. [1]

Both the access and disambiguation pruning algorithms work by using probability to rank constructions to be accessed or interpretations to be disambiguated. In the next section we introduce the grammatical formalism that will allow these probabilities to be computed.

2.3 Grammar: Uniformity and Probability

The representational component of our processing model is based on construction grammar. Construction grammar adopts from traditional grammar the idea that a grammar consists of a large collection of grammatical constructions. Each grammatical construction is a sign, a conventionalized association between form and meaning. Formally, these signs are represented as typed unification-based augmented context-free rules in the version of construction grammar we consider here. In these senses

[1] See Mathis and Mozer (1995) for other arguments for the computational utility of disambiguation.

construction grammar resembles most phrase-structure and unification-based linguistic theories, including among others LFG (Bresnan 1982), HPSG (Pollard and Sag 1987) [2], and certain versions of Categorial Grammar. Other versions of construction grammar, and closely related theories like Cognitive Grammar (Langacker 1987), share the sign-based foundation of the version of construction grammar we discuss here, but differ on the use of feature structures and CF rules as a formal implementation. Like all these similar approaches, construction grammar contrasts with the non-sign-based Principles and Parameters/Minimalist approach to grammatical representation. This paper will also show how to augment construction grammar with certain kinds of probabilities. As we will see below, these same probabilistic augmentations could be made to any other of these sign-based theories which meets certain requirements.

Our processing model relies on four assumptions concerning grammatical representation:

1. The representation of constituent structure rules as mental objects.
2. A uniform context-free model of lexical, morphological and syntactic rules.
3. Valence expectations on lexical heads.
4. A lack of empty categories.

In the rest of this section we describe and motivate each of these assumptions, discussing the extent to which each is true of construction grammar and other theories.

The first assumption deals with what Fillmore et al. (1988) refer to as "the distinction between knowing and figuring out", what is often called in psychology the 'economy of representation' hypothesis. Henderson (1989) has remarked that linguistics and psychology hold quite different views about representational economy. A fundamental motivating principle for many linguistic theories is the minimal-redundancy, maximally economical grammar or lexicon. Psychological models, on the other hand, often emphasize the vast storage capability of the mind. Construction grammar borrows from these tendencies in psychology by modeling grammar as a comparatively large collection of structures, including larger phrase-structure rules as well as lexical entries. Generalizations across these structures, including simple lexical entries as well as complex phrase-structural projections such as the correlative conditional construction of Fillmore (1986) described below, are captured by a type hierarchy, resembling the type hierarchy of HPSG or the schematic networks of cognitive grammar. Figure 1 shows part of the construction grammar type hierarchy.

The idea that the grammar is a collection of these structures arranged in a type hierarchy has two implications for processing. First, because constructions and the type hierarchy are the sole mechanism for capturing generalizations, the theory places no requirement on the processing mechanism to implement redundancy rules, metarules, movement, or enrichment of surface form.[3]. We will use the lack of traces to model

[2]Note that uniform processing of lexical and syntactic knowledge can take place whether we assume that lexical entries or CF rules are the more fundamental mechanism; i.e. lexical entries can be modeled as trivial lexical insertion rules in a CFG formalism, or alternatively phrase structure constituency can be represented with special attributes with complex feature structures.

[3]For arguments that inheritance together with on-line type construction is sufficient to replace lexical rules see Koenig and Jurafsky (1995).

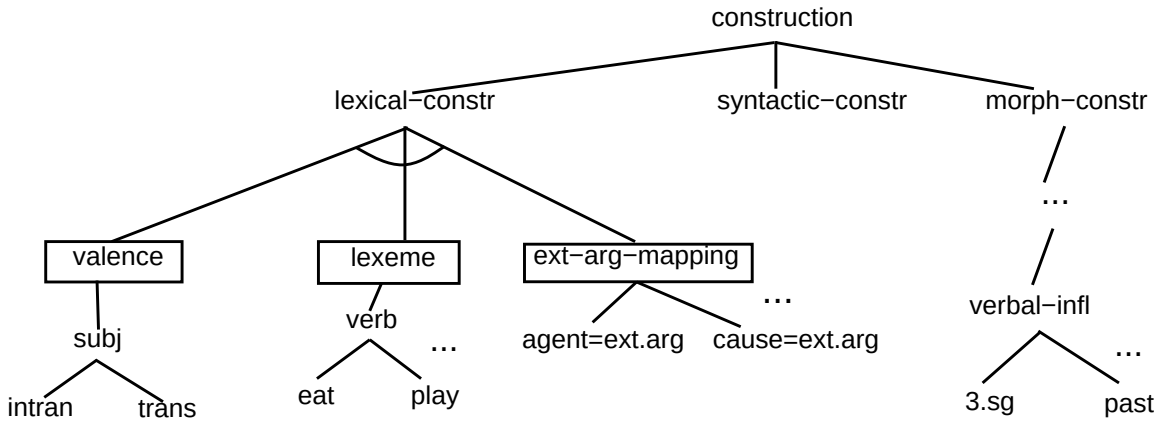


Figure 1: Part of the Construction Grammar Type Hierarchy (from Koenig & Jurafsky (1995)

the psycholinguistic results of Tanenhaus et al. (1985) in §4. In as much as LFG has eliminated traces, and recent versions of HPSG have begun to do away with traces (Pollard and Sag 1994) and lexical rules (Krieger and Nerbonne 1993), this assumption of representational minimalism is met by a number of theories. Second, allowing these structures means the theory has an independent, relatively rich phrase-structural component, in which phrase-structural patterns are stored in the mental inventory. We will see the need for storing constituent-structure patterns (and their probabilities) in modeling garden-path effects in §4. Besides construction grammar, the storage of these structures is assumed by the sign-based theories (HPSG, Cognitive Grammar) and LFG, but not by categorial grammar, and also not in the PP/Minimalism paradigm, in which phrase structural categories act only as a kind of principle-based filter on possible trees, and have no independent existence in any kind of mental store (and hence cannot be augmented with probabilities).

The second fundamental assumption of construction grammar is representational uniformity. Uniformity has a number of facets. First, construction grammar makes no distinction between the lexicon, the morphology and the syntactic rule base; each is represented with an augmented CFG formalism. A lexical entry is a construction that may have only a single constituent. Morphological knowledge is represented with context-free rules (Selkirk 1982) for example an affix like English plural 's' may be treated as a rule *N[+plural]* $\rightarrow$ STEM *'s'*.[4] This assumption allows our processing model to treat lexical and syntactic processing equivalently; although we don't consider morphology in this paper, we expect to apply the same parsing algorithms for morphology as we have for syntax. The idea that lexical, morphological, and syntactic structures are uniformly represented is a fundamental principle of sign-based linguistic theories such as Cognitive Grammar and categorial grammar. With the exception of morphological knowledge this is also true of HPSG.[5] Because in sign-based theories

[4]See Langacker (1987) for a related, although non-context-free, proposal within Cognitive Grammar and Koenig (1994) and Orgun et al. (1995) for a development of context-free morphology within construction grammar

[5]Although LFG makes a crucial distinction between the lexicon and the grammar, recent LFG work has discussed similarities between lexical and syntactic structure such as assigning semantic rules to specific syntactic constructions

like construction grammar constructions can be associated with semantics like lexical items, the access and disambiguation processes can make reference to semantic structure.

As an example of a construction and the use of inheritance to capture generalizations, consider the CORRELATIVE CONDITIONAL construction described by Fillmore (1986) & (1988) and McCawley (1989), which models sentences like (2):

(2) a. *HERMIA*: The more I hate, the more he follows me.

b. *HELENA*: The more I love, the more he hateth me.

Figure 2 shows Fillmore's analysis of the construction.

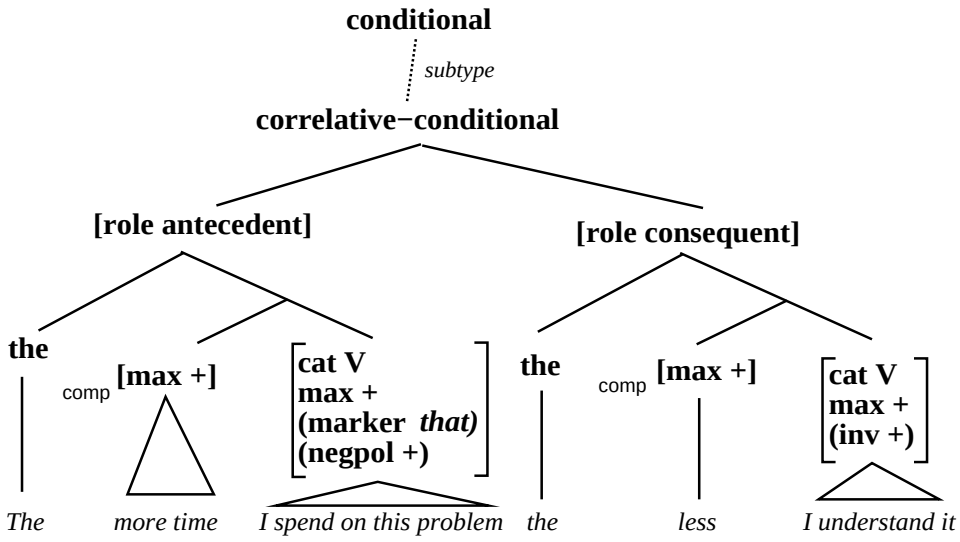


Figure 2: The Correlative Conditional Construction (after Fillmore (1988))

Note that the correlative conditional construction is a subtype of the conditional construction, which is itself a subtype of the subordination construction. In addition, the second and fifth constituents are constrained to be of type COMPARATIVE. Fillmore argues that sentences like (2) simultaneously exhibit conditional and comparative properties, and that these various grammatical properties of the construction are inherited from the conditional and comparative constructions.

For example the correlative conditional inherits from the conditional construction many facts about the tense and aspect possibilities of the two clauses, such as the suppression of future *will* in the protasis:

(3) a. The faster you (*will) drive, the sooner you'll get there.

b. If you (*will) drive fast, you'll get there by 2:00.

As McCawley (1989) shows in (4) , the second and fifth constituents of the correlative conditional exhibit the complete range of ordinary comparative morphology; *-er* on short stems, *more* on long stems, and suppletion with *good, bad, will, badly, much, many, little*.

like relative clauses (Dalrymple 1992).

(4) a. The worse the weather gets, the happier I am that we stayed home.
b. The worse he behaved, the less attention we paid to him.
c. The better you treat him, the less trouble he'll give you.

The correlative conditional has a number of idiosyncratic properties which cannot be predicted from its supertypes. For example, some particulars about the use of the comparative are unique to this construction; the two comparatives ("the more time" and "the less") are fronted and semantically express a correlation between an independent and a dependent variable. For a sentence like (5), the interpretation might be paraphrased as "The degree to which I spend time on this problem will determine the degree to which I understand it".

(5) The more time I spend on this problem, the less I understand it

Again, a guiding intuition of construction grammar is that such complex, non-local constructions are stored along with lexical entries in the mental grammar.

We turn now to the probabilistic augmentations we propose for constructions. We argue that constructions should be augmented with probabilities in two ways. First, every construction will be annotated with a simple probability. These express the prior probability of choosing a particular construction at random from the mental lexicon or from a random utterance. This prior probability resembles the "resting activation" of the frequency-based models traditional in psychological theories of lexical access. Second, constructions are augmented with probabilities in every case where some part of the construction expresses a linguistic expectation. A feature-based lexical grammar expresses expectations in at least two ways. First, each context-free rule expresses an expectation for the constituents which constitute the rule. That is, given that we have evidence that the parser has seen a certain phrase structure rule, we can expect each constituent of the rule to appear. Second, each valence-bearing predicate expresses an expectation for its valence-fillers. That is, given evidence for a predicate, we can expect to find further in the string a constituent which will satisfy each valence (or subcategorization) slot of the predicate. Our model allows each of these expectations to be probabilistic.

How are these probabilities to be defined? The prior probability of a construction is the simplest. We assume here a frequentist interpretation of prior probabilities. That is, the prior probability of a construction is just a maximum likelihood estimate from relative frequencies. We make the simplifying assumption that these frequencies can be computed either from psychological norming studies or from carefully chosen corpora. The most widely available balanced corpus is the million word Brown corpus. Francis and Kučera (1982) give frequencies for every lexical item in the Brown corpus, and also for simple lexical categories. Computing frequencies for phrasal constructions requires parsing the corpus. For phrasal construction frequencies we used the Penn Treebank (Marcus et al. 1993), which includes a parsed segment of the Brown corpus.

Given the prior probabilities for each construction, computing the probabilities for constituent expectations can be done with a simple normalization step. Constituent expectation probabilities are in fact the standard Stochastic Context-Free Grammar probability. This is the conditional probability of the right hand side of a rule, given

the left hand side, i.e. the probability of a particular expansion of a left-hand side. Thus if the expansions of A consist of SCFG rules of the form

(6) $[p_1]$ *A* → *B C*

(7) $[p_2]$ *A* → *D E F*

then p_1 is the probability that A will expand to BC, while p_2 is the probability that A will expand to DEF. Since these probabilities are all conditional on A, it follows that $\sum_i p_i = 1.0$. Thus we can compute the SCFG probability of any rule just by dividing the prior probability of the rule by the sum of the prior probabilities of all expansions of the left-hand side.

For example, in order to compute the probability of the simplified rule (8), we would use the treebank to get a frequency for all NPs (52,627), and then for those NP's which consist of a Det and an N (33,108). The conditional probability is then 33,108/52,627=.63.

(8) [.63] *NP* → *Det N*

The second kind of grammatical expectations are valence expectations that a lexical predicate bears for its arguments. As we discussed above, the valence of a predicate in construction grammar may constrain the syntactic category, grammatical function, thematic role, or semantics of their arguments. Each lexical predicate may bear a valence attribute which specifies a set of valence arguments. In our probabilistic model, each of these arguments is augmented with a probability of occurrence. This probability expresses the conditional probability of the argument being filled given the predicate. Obligatory arguments, for example, will have unity valence probability, while optional arguments will have a value between 0 and 1.

An alternative algorithm might assign probabilities to entire thematic grids rather than individual arguments. Although we express our probability computations in terms of individual-argument rather than thematic-grid probabilities, we will show our pruning algorithm is compatible with either model. The data we discuss in this paper does not allow us to definitively choose between the two models of valence probability.

We have relied on two methods of determining valence probabilities. Connine et al. (1984) provide frequencies for the different syntactic subcategorizations of 127 verbs. These were determined by giving subjects a list of verbs and asking them to write a sentence for each one. The authors then computed frequencies on the corpus this produced.

For verbs which were not studied by Connine et al., we used the Penn Treebank (Marcus et al. 1993) just as for construction probabilities. That is, to determine the valence probabilities for a verb, we counted the number of times the verb occurred with each frame, and normalized by the total count for the verb. Since these initial sources of valence probabilities only include syntactic information, we have left the use of semantic valence probabilities for future research (see §5).

3 Access

Every model of parsing, lexical access, or sentence processing includes a theory of *access*, and yet the problem of access has been strikingly balkanized. In this section

we compare previous models of access at the lexical, idiomatic, and syntactic levels, and hold them up to psycholinguistic results. We show that each of the models fails to model all the relevant data, and then present the Bayesian access model, based on computing the conditional probability of a construction given top-down and bottom-up evidence. We show that the model is consistent with psycholinguistic results and outline the model's predictions.

3.1 Previous Models of Access

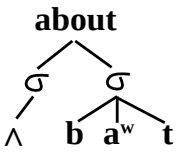
Consider the difference between *serial* and *parallel* access algorithms. The earliest models of lexical, idiomatic, and syntactic access were all serial, including lexical models like *ordered-access* (Hogaboam and Perfetti 1975), models of idiom access like the idiom list hypothesis (Bobrow and Bell 1973) and the direct access hypothesis (Gibbs 1984), and models of syntactic access in parsers like PARSIFAL (Marcus 1980) and the Sausage Machine (Frazier and Fodor 1978). In serial models a single lexical entry is accessed from the lexicon at a time. For example, in the idiom list hypothesis, idioms are stored in a separate idiom dictionary, and a single idiom is accessed when the computation of literal meanings for a string fails, while the direct access model of idiom access proposes just the opposite: idiomatic meaning is accessed first, and literal meaning is only computed when the access of idiomatic meaning fails.

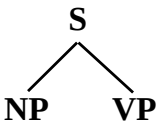
In recent years, serial models of lexical and idiom access have fallen somewhat out of favor, due to extensive results which argue for parallelism (Swinney 1979; Swinney and Cutler 1979;etc). Thus most researchers on idiom processing assume some form of the lexicalization hypothesis (Swinney and Cutler 1979), which proposes that idioms are stored as long lexical items and accessed in parallel with normal lexical items.

Syntactic access is treated very differently in the psycholinguistic and computational communities. Psycholinguistic parsers (such as the early models of Bever (1970) and Kimball (1973), as well as later models like the Sausage Machine), are traditionally serial, motivated as they are by the garden-path results we will discuss in §4. Computational parsing models (such as the Earley, CKY, and chart parsers) are usually based on some form of dynamic programming algorithm, and hence are fundamentally parallel in nature.

However, we will argue in §4 following Kurtzman (1985), Norvig (1988), and Gibson (1991), that garden-path phenomena can be modeled with a parallel architecture as well, eliminating one of the strongest arguments for serial parsing. Although psycholinguistic results on syntactic access seem to be somewhat contended, a number of recent results argue for parallelism in syntactic processing as well (Kurtzman 1985; Gorrell 1987; Gorrell 1989; Boland 1991; MacDonald 1993). Because all modern models of lexical access, and most models of idiom or syntactic access are parallel, then, the remainder of this section focuses on parallel access algorithms, examining previous models from two perspectives: *how* to access (i.e. what sorts of evidence to use), and *when* to access (i.e. the time course of access). For example, consider the problem of accessing either the syntactic rule $S \rightarrow NP\ VP$ or the lexical item 'about' ($\Lambda\ b\ \mathfrak{a}^w\ t$) given some input. Figure 3 sketches four standard rule access models, showing the type of evidence each algorithm would use to access the rule $S \rightarrow NP\ VP$

or 'about' → ʌ *b* ɑʷ *t*.





Access Method	Examples	Phonological Evidence	Syntactic Evidence
• **Bottom-Up**	shift-reduce	ʌ *b* ɑʷ *t*	NP VP
• **Top-Down**	LL(k), selective access	about	S
• **Left-Corner**	Cohort	ʌ	NP
• **Island**	head-corner, key, MSS	*b* ɑʷ *t*	N

Figure 3: Previous Models of Phonological and Syntactic Access

In the *bottom-up* model, access of a rule depends only on evidence from the input, and is delayed until the entire set of constituents has been seen. One interpretation of the lexicalization hypothesis (Swinney and Cutler 1979) would process idioms in this way; van der Linden and Kraaij (1990) built a computational model of this proposal, in which idioms are not accessed until the entire idiom has been seen. In a top-down model (such as the *selective access* model of lexical access (Schvaneveldt et al. 1976; Glucksberg et al. 1986)), access depends only on previous expectations. Neither the top-down nor bottom-up models meets our concern with psychological plausibility. For example, we cited in the introduction a large number of studies showing that language processing is strictly on-line, ruling out a bottom-up model which delays until every constituent has been seen. Similarly, a number of studies have shown results inconsistent with selective access and other purely top-down models (Swinney 1979).

The *left-corner model* combines advantages of the bottom-up and top-down models; a rule is accessed by only the first constituent, and then processing continues in a top-down manner from the rest of the rule. Such models have been proposed for both syntactic parsing and lexical access. For example, Marslen-Wilson's (1987) Cohort model of lexical access in speech perception is in many respects a left-corner model, using bottom-up information to access entries, and then top-down information to process them further. In the Cohort model, bottom-up phonetic information is used to access a set of lexical entries whose initial phonemes are compatible with the input so far. The set, called the *cohort* set, is then weeded out by using various top-down as well as bottom-up information sources to prune words which don't meet their constraints.

The final class of models, the *island* models, propose even more sophisticated ways of accessing a rule. In *head-corner* access, only the *head* of the first constituent need be seen to access a rule. While the head-corner model was proposed independently by van Noord (1991) and Gibson (1991) for syntactic parsing, Cutler and Norris's (1988) MSS model of lexical segmentation is essentially a head-corner model applied to speech. The MSS model accesses words based on their first stressed syllable; the stressed syllable thus acts as the parsing head.

Finally, in what can be viewed as an extension to the head-corner model, two algorithms (Wilensky and Arens 1980; Cacciari and Tabossi 1988) have been proposed which mark specific constituents of idioms as the *key* or *indexing clue*, and access idioms only after this constituent is seen. This allows these algorithms to model results indicating that the access of different idioms will occur at differing positions in the idiom.

3.2 Problems with Previous Models

Clearly there is a trend in more recent access models to be more and more sophisticated about the kind of evidence that is needed to access a rule or structure. Unfortunately none of these models is quite sophisticated enough, and all suffer from two major problems. The first is their inability to handle *timing* effects; in particular construction-specific, context-specific, and frequency effects in access. The second is their reliance on a single kind of information to access rules; either strictly bottom-up or top-down information, strictly syntactic information like the restriction algorithm of Shieber (1985), or solely semantic information, as in conceptual analyzers like Riesbeck and Schank (1978) or in Cardie and Lehnert (1991) or Lytinen (1991). Thus each of these models is only able to model a particular range of evidential effects.

Consider the psycholinguistic evidence on timing. First, there is evidence that access timing is different for different constructions. The *access point* (point in time when the construction is accessed) for different constructions may be quite distinct. For lexical constructions, Tyler (1984) and Salasoo and Pisoni (1985) show that while the average access point for lexical items is approximately 150 ms after word-onset, timing is quite dependent on the frequency of the lexical item. High-frequency lexical items have higher initial activation than low-frequency ones (Marslen-Wilson (1990)), are accessed more easily (Tyler 1984 and Zwitserlood 1989), and reach recognition threshold more quickly (Simpson and Burgess 1985 and Salasoo and Pisoni 1985). Swinney and Cutler (1979) showed that some idioms were not accessed immediately after the first content word of the idiom, but rather that it took at least two words to access the idiom. Cacciari and Tabossi (1988) found that different idioms are accessed at quite different rates.

One way to model the access timing differences between constructions is with the *island* or *key* algorithm of Cacciari and Tabossi (1988) and Wilensky and Arens (1980). In these algorithms, each construction in the grammar would have one of its constituents (the *key* or *island*) marked. The construction could only be accessed after the key had been seen. One problem with this approach is that it requires specifying this information for each construction in the language; an approach which was able to predict the access point without having to learn it for each construction would be preferable.

A more serious problem with the island/key algorithm is its inability to account for the second class of timing results. These results show that different contexts change the access point even for the same construction, i.e., that the access point is context-sensitive. Cacciari and Tabossi (1988) showed that the access of idioms was faster in the presence of context. Salasoo and Pisoni (1985) showed the same for lexical entries.

Marslen-Wilson et al. (1988) showed the dual case — that anomalous contexts can slow down the access point of lexical constructions, and that the more anomalous the contexts, the higher the response latencies.

Thus, whatever sort of access algorithm we propose, it must allow the accumulation of evidence to cause some constructions, in some contexts, to be accessed faster than other constructions, in other contexts.

3.3 The Probabilistic Model

Our proposal is that access and disambiguation should be treated with a single probabilistical model. We refer to the access part of the model as the *Bayesian* or *evidential* access algorithm. For each construction, we compute the *conditional probability* of the construction given the evidence. Evidence can come from syntactic, semantic, and lexical sources, both top-down and bottom-up. Constructions are accessed according to a *beam-search* algorithm. In beam-search, every construction falling within a certain percentage of the most highly-ranked construction is accessed. We propose that this beam-width, which we call the *access threshold* α, is a universal constant in the grammar.

For each construction, the conditional probability of a construction given top-down evidence is relatively simple to compute in a Construction Grammar or any other augmented-stochastic-context-free formalism. Recall that the SCFG prior probability gives the conditional probability of the right hand side of a rule given the left hand side.

Given these probabilities, the conditional probability of a construction c appearing given some top-down evidence construction e^+ can be computed from the probability that e^+ will expand to c. In particular, since the parser operates left to right, the top-down probability $P(c|e)$ is the probability that the evidence construction e^+ left-expands to c:

(9)
$$P(e \stackrel{L*}{\Rightarrow} c)$$

In a context-free grammar, a nonterminal a left-expands to a nonterminal b if there is some derivation tree whose root is a and whose leftmost leaf is b. Consider the toy example in (10).

(10) a. [.5] *S* → *NP VP*
b. [.7] *VP* → *V NP*
c. [.2] *VP* → *V NP PP*
d. [.1] *VP* → *Adv VP*
e. [.1] *V* → *'eat'*
f. [.3] *V* → *'go'*

Suppose the parser is in the process of parsing a sentence and has found the first NP of an S. The left expansion of the next constituent, the VP, includes the symbols V, eat, Adv, and go. Figure 4 shows a schematic example of left expansion.

Equation 9 holds whether the evidence is *valence evidence*, i.e. evidence from a lexical head that a particular complement (syntactic or semantic) is expected, or

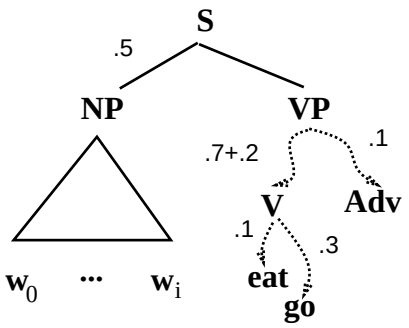


Figure 4: Left-Expansion: a schematic example

constituent evidence, i.e. evidence from an incomplete construction. If no recursive production cycles occur in the expansion of e to c, then $P(c|e)$ is very simple to compute as the *product* of the probabilities associated with each rule in every possible expansion. If recursive production cycles occur, Jelinek and Lafferty (1991) give an algorithm for efficiently computing left-corner probabilities, while Stolcke (1995) shows how the standard Earley parser can be augmented with them. Jurafsky et al. (1995)b shows an application of a version of this algorithm as a model of lexical access for a speech recognizer.

In order to compute the conditional probability of a construction c given evidence e which is *bottom-up* evidence, we use the Bayes Rule:

(11)
$$P(c \mid e) = \frac{P(e \mid c)P(c)}{P(e)}$$

The probabilities $P(c)$ and $P(e)$ are simply the prior probabilities which annotate each construction, computed by normalizing frequencies. For the likelihood $P(e \mid c)$, we can now use the standard algorithm above in reverse, computing the probability that c left-expands to e.

Once given these estimates for top-down and bottom-up probabilities, combining the evidence to estimate the posterior probability of a construction is quite complex. In current work, we are examining Bayes net realizations of the combination problem, which is in fact the subject of a large Bayes Net literature (Pearl 1988). Certain assumptions may make the combination problem much simpler. For example, if we assume that top-down and bottom-up evidence interact disjunctively, they can be combined with a simple noisy OR-gate. Disjunctive interaction occurs when any of a set of conditions is likely to cause a given event and this likelihood does not decrease when more than one of the conditions hold. This assumes that if a construction c is likely given either one of two causal conditions e_1 and e_2, then each condition has an independent chance of causing the construction, and also assumes exception independence, i.e. that whatever inhibits e_1 from causing c is independent of whatever inhibits e_2 from causing c. Indeed, these are probably weaker assumptions than the assumption of rule-independence implicit in stochastic context-free grammars. The noisy OR-gate multiplies the probability of exceptions to compute the exception probability for c; if e^+ is the top-down evidence for a construction and e^- is the bottom-up evidence, the probability of the construction is

(12) $$P(c|e) = 1 - (1 - P(c|e^+))(1 - P(c|e^-))$$

Alternatively, we can make the simplifying assumption that e^+ can effect e^- only through c, producing the following:

(13) $$\begin{aligned} P(c|e) = P(c|e^+, e^-) &= \frac{P(c, e^-|e^+)}{P(e^-|e^+)} \\ &= \frac{P(c|e^+)P(e^-|c, e^+)}{P(e^-|e^+)} \\ &= \frac{P(c|e^+)P(e^-|c)}{P(e^-|e^+)} \end{aligned}$$

Since, as we show below, we will be comparing the *ratios* of the probabilities for different constructions, the denominator $P(e^-|e^+)$ remains constant and we can simply treat it as a normalizing factor, producing a very simple equation for evidence combination:

(14) $$P(c|e) = \alpha P(c|e^+)P(e^-|c)$$

However, psycholinguistic results on the relation between top-down and bottom-up evidence are quite controversial. While most studies agree that top-down evidence is used in parsing, many researchers have argued that bottom-up evidence is used first or more strongly (Mitchell 1989; Clifton and Frazier 1989) (although cf MacDonald 1993). This might be modeled by parameterizing the amount of weight given to top-down versus bottom-up evidence, perhaps via exponential weighting of the probabilities. A more complete investigation of evidence combination awaits future work.

Now given that the algorithm assigns a conditional probability to each construction, how should we choose which ones to access? We might choose to access any construction with non-zero probability. But as we argued in §2.2, accessing every possible construction may lead to far too many constructions, especially with noisy input such as speech. Furthermore, the results of Cacciari and Tabossi (1988) discussed above argue that idioms are often not accessed until more than one word of them has been seen. This suggests that the access algorithm operates with some sort of pruning. For example, if the phonetic evidence suggests both a lexical entry and an idiom, but the lexical entry was much more probable, a pruning algorithm might keep the idiom from being accessed. Since the data rules out any sort of absolute threshold, our proposal for both access and disambiguation pruning is to use relative-width beam search, which prunes any construction more than a constant times worse than the best construction. If an idiom is much less probable than a lexical item, it would only be accessed if enough of the idiom had been seen to provide more evidence for the idiom than just for the lexical item. Our preliminary proposal is that there is a single fixed access beamwidth threshold for the grammar, α, and that all constructions are accessed whose probability is within a multiple of α of the most likely construction. Because there is insufficient psycholinguistic data on syntactic access in particular, we do not propose a specific value for α in this paper. We do, however, propose a specific range of values for the disambiguation beam-width in §4.

Let's work through one example of access. Suppose we are parsing a sentence beginning '*who can . . . '*. The parser has just seen the word 'who', and already accessed a couple of the sentential constructions which model wh-questions, the MAIN-CLAUSE-NON-SUBJECT-WH-QUESTION and the SUBJECT-MATRIX-WH-QUESTION. Both of these constructions are marked with *role [gf main]* to indicate that they are main-clause-level constructions, and *syn [vif fin]*, indicating that they will have finite verb phrases. Both of these constructions begin with a wh-element, marked *[syn [wh +]]* in Figure 5. In the SUBJECT-M-WH-QUESTION this is followed by a VP (a fact which is inherited from the SUBJECT-PREDICATE construction; in the MC-NON-SUBJECT-WH-QUESTION this is followed by an inverted clause.

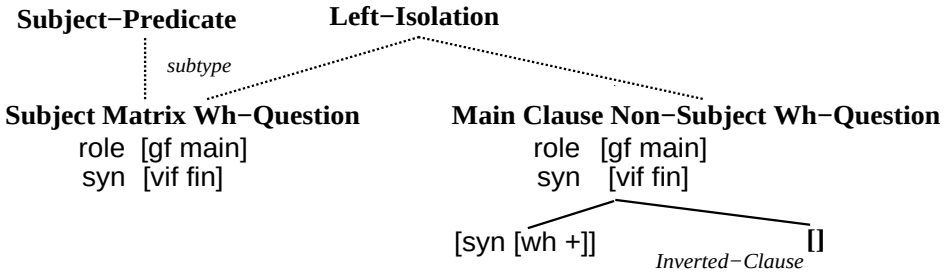


Figure 5: A part of the Construction Grammar type hierarchy

The inverted clause construction ('**are you there**?', 'so **will she**') requires an Aux as its first constituent. The VP construction requires any verb as its first constituent. Consider now the dynamic programming chart on which the parser will keep records of interpretations in progress. The chart will have an arc for each construction being considered. Assume for simplicity that these were the only two constructions active on the chart, as shown in Figure 6. Each arc on the chart is associated with a construction, which has a dot to indicate the current place in the parse. Thus both constructions have already seen their first constituent (marked *[syn [wh +]]*).

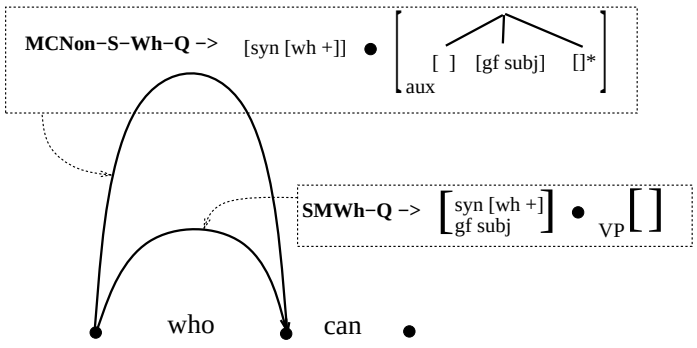


Figure 6: Top-down expectations from two arcs on the chart

The probability based on top-down evidence for an Aux appearing next given the MC-Non-Subject-Wh-Question is 1.0, since the construction requires an Aux to appear. Since AUX subsumes CAN in the type hierarchy, the likelihood P(CAN|AUX) is simply

P(CAN)/P(AUX). Computing the prior probabilities of each of these construction from the Brown Corpus (Francis and Kučera 1982), gives us P(CAN)/P(AUX) = .0017/.025 = .07. The probability based on top-down evidence for any verb appearing next given the Subject-M-Wh-Question is 1.0, again since a verb is required. Since the prior probability of a VERB in the Brown Corpus is .085, this makes P(CAN|SUBJECT-M-WH-QUESTION)= P(CAN|VERB)= P(CAN)/P(VERB)=.0017/.085 = .021. The probabilities of 'can' given the two Wh-Question constructions are independent, and so the actual probability based on top-down evidence of 'can' is a weighted sum:

(15) $$P(can|e^+) = P(\textit{MC-Non-S-Wh-Q}|\textit{context})P(\textit{can}|\textit{MC-Non-S-Wh-Q}) + P(\textit{S-M-Wh-Q}|\textit{context})P(\textit{can}|\textit{S-M-Wh-Q})$$

$P(\text{MC-Non-Subject-Wh-Question}|\text{context})$, assuming just the context of this sentence, i.e. the first word 'who', is the probability of the derivation of this prefix 'who' from the MC-Non-Subject-Wh-Question. Whatever these probabilities are, the final $P(can|e^+)$ will be between .021 and .07. This is the evidence that will be combined with any bottom-up evidence for 'can'.

3.4 Advantages of the Model

The probabilistic model of access accounts for a number of psycholinguistic results. First, the model explains the frequency effect in lexical access. Lexical items with a higher frequency will tend to have a higher probability. We can see this by considering the equation for bottom-up evidence, applied to lexical access:

(16) $$P(w \mid e) = \frac{P(e \mid w)P(w)}{P(e)}$$

(16) implies (17), i.e. that the probability of any word w is directly proportional to its prior probability $P(w)$, which we take from maximum likelihood estimates from relative frequencies.

(17) $$P(w \mid e) \propto P(w)$$

Since idioms and lexical items are treated equivalently, the model thus also accounts for experiments on idioms which show that more familiar idioms are accessed more quickly (d'Arcais 1993). Finally, the model predicts that this frequency effect should extend to syntactic constructions.

But the probabilistic algorithm differs from a simple frequency-based algorithm in making the prediction that access of a construction will be *inversely* proportional to the probability of the evidence. In other words, the relationship between the posterior probability of a construction and the priors of evidence and construction is as follows:

(18) $$P(c \mid e) \propto \frac{P(c)}{P(e)}$$

Intuitively, things which are common are not good evidence for any one construction in particular. This aspect of using posteriors rather than priors to model access explains experimental results in both lexical and idiom processing.

First, the model provides a way to formalize the intuitions of Cacciari and Tabossi's (1988) results on idiom access. Cacciari and Tabossi distinguished predictable and unpredictable idioms in Italian. Predictable idioms are those like *in seventh heaven* in which they considered there to be an early cue to idiomaticity. Unpredictable idioms were those like *go to the devil* in which the cue that the phrase was not literal came late in the idiom. They found that the idiomatic meaning was accessed immediately for predictable idioms, but much later for unpredictable idioms (300 ms after offset). Since the authors controlled for frequency of idioms, this cannot be accounted for by assuming that more frequent idioms were accessed earlier.

But consider what it means to be a cue to idiomaticity. Essentially the longer the structure of the idiom resembles a literal phrase, the later the cue. That is, the cue is marked by a particularly unlikely or rare word, combination of words, or construction. This is exactly what we would expect from a probabilistic approach. A more common structure, i.e. one which is a very likely candidate for literal interpretation, will have a high probability. A further examination of their late-cue idioms showed that many of them also began with very common words such as *venire ('come')*, or *andare ('go')*. The probabilistic interpretation of this result, then, is that the highly probable words or constructions which began the idiom did not prove a good source of evidence for the idiom, because they provided evidence for so many other constructions as well. Again, this supports the use of a true probabilistic model, rather than the simple relative-frequency model assumed by previous lexical access theories.

Next, the probabilistic model explains the similarity neighborhood effects found by (Luce et al. 1990) and others. Luce et al. showed that words differ in their *similarity neighborhoods*, the set of words which resemble them phonetically. They showed a correlation between the time to access a word and the size of its neighborhood; words with large neighborhoods were slower on auditory naming and lexical decision. These effects are predicted by the probabilistic model just as for idiom access. The difference between a frequency-based and a probabilistic theory of lexical access are the likelihood term $P(e|c)$ and the prior of the evidence $P(e)$ in (16). The likelihood term will model the goodness-of-fit of a lexical entry to the input; we don't consider this further. If the evidence for a word is some (perhaps noisy) string of phones, then the evidence prior expresses how common this string of phones is a priori, perhaps the phonotactics of the language. But consider what it means to have a large frequency-weighted neighborhood; it means that there are a lot of words which strongly resemble the target word. But if a lot of words resemble the target word, the target word must have a very common phonology. In other words, a large neighborhood means the prior probability of the evidence will be high. A small neighborhood means there are very few words in the language whose phonotactics resemble the target word, and hence the string of phones used as evidence will have a small probability.

Luce et al. (1990) explained the neighborhood effects by adding terms for neighborhood size and frequency to R.D. Luce's (1959) choice rule. We show that their modified Luce rule is an approximation to our probabilistic model. Luce et al. (1990:125) gives the following equation:

(19)

$$P(ID) = \frac{P(\text{stimulus word}) \times freq_s}{P(\text{stimulus word}) \times freq_s + \sum_{j=1}^{n} p(\text{neighbor}_j) \times freq_j}$$

What they describe in their equation as $P(\text{stimulus word})$ corresponds to our likelihood term, but their method of computation blurs the distinction between likelihood and posterior. Using confusion matrices between underlying and surface phonological form, they compute $P(\text{stimulus word})$ as $P(\text{stimulus word}|\text{input string})$ rather than $P(\text{input string}|\text{stimulus word})$. For example, they compute the probability of the stimulus word */kæt/* by multiplying together the confusion probabilities *P(k| k), P(æ|æ),* and *P(t | t)*. Since with no deletions, insertions, or reductions, these probabilities are symmetric, their computation does in fact simulate the actual likelihood computation. But the actual likelihood computation would be completely different if the lexical entry for a word does not have the same string of phones as the surface string (e.g. for deletions, $P(\emptyset|t) \neq P(t|\emptyset)$).

The denominator of (19) incorporates the neighborhood size and confusability with the target word. But this equation can also be viewed as a heuristic to approximate the actual prior probability of the evidence. By definition, the true prior probability of the evidence is equal to a weighted sum over all words l in the lexicon of the conditional probability of the evidence given l:

(20)
$$P(e) = \sum_{l \in lexicon} p(e|l)p(l)$$

What Luce et al. (1990) compute only approximates this true prior. First, they again compute the posterior P(neighbor word | input string) rather than P(input string | neighbor word). As we discussed above for the likelihood computation, the difference is not very great in their case, since they use single-phone confusion matrices and a simple string-of-phones model of phonology. But again, the more different the neighbor word is from the target, the more likely that asymmetries in the confusion matrices will cause their equations to diverge from the true probabilities. Second, they only consider the probability of neighbor words, rather than all words in the lexicon. In practice, this is probably a good approximation, since it is the neighbor words that will provide the bulk of the probability mass, since the likelihood $p(e|l)$ will be extremely low for words which are very different than the target word. However, it is only an approximation, and requires making an arbitrary decision about what constitutes a neighbor. Computing the true prior requires no such arbitrary definitions of neighborhood.

A final source of evidence for the Bayesian evidential theory has to do with the varying use of top-down evidence in syntax versus phonology. Tanenhaus and Lucas (1987) note that psycholinguistic evidence of top-down effects is very common in phonology, but much rarer in syntax. Researchers have noted that top-down syntactic/semantic effects on lexical access are only reported in extremely strong contexts, particularly with lexical contexts. For example Wright and Garrett (1984) found evidence for top-down syntactic effects by showing that very strong syntactic contexts affected the reaction time for lexical decisions on nouns, verbs, and adjectives. In one

experiment, a context ending in a modal verb sharply reduced the time for lexical decision to a verb. Similarly, a context ending in a preposition reduced the time for lexical decision to a noun. In phonology, the conditional probability of a phoneme appearing given a word in which it occurs is very high, and thus top-down evidence will be quite high. Syntactic constraints, on the other hand, are generally specified in terms of very abstract constructions like NOUN or VERB. Thus the top-down conditional probability of any *particular* noun appearing is quite low, and top-down evidence will be much lower.

The Bayesian access theory predicts that top-down syntactic effects are more likely in cases where a syntactic construction refers to a particular lexical item. Here the conditional probability of this given word occurring will be quite high. Preliminary evidence for this prediction comes from Cacciari and Tabossi (1988), who found just such top-down syntactic effects from idioms.

We can summarize the psycholinguistic data which support our probabilistic algorithm as follows:

- The access-point of a construction varies across constructions and contexts.
- Evidence for a construction is weighted in direct proportion to the frequency of the construction.
  - Lexical access time correlates directly with word frequency.
- Evidence for a construction is weighted in inverse proportion to the frequency of the evidence.
  - Lexical items with high frequency-weighted neighborhoods take longer to access.
  - Idioms with high-frequency words take longer to access.
- Top-down evidence for a construction is weighted in proportion to the relative specificity of the evidence and the construction.

In the next section we continue our arguments for a probabilistic model of language processing, showing that using probabilities to rank interpretations of an ambiguous utterance, just as we rank constructions to access, can account for psycholinguistic results on disambiguation.

4 Disambiguation

Natural language is inherently ambiguous. In studying the response of human subjects to this ambiguity, cognitive studies have amassed a large collection of results to be accounted for by a computational model. This includes on-line and off-line experiments on parsing preferences, (Frazier and Rayner 1987; Britt et al. 1992; Ford et al. 1982; Crain and Steedman 1985; Whittemore et al. 1990; Taraban and McClelland 1988) results on the interpretation of garden-path sentences, (Bever 1970; Kurtzman 1985; Frazier and Rayner 1987) and studies of gap-filling/valence ambiguities (Tanenhaus et al. 1985, etc).

Whether a model of parsing is serial or parallel, disambiguation preferences are accounted for by applying a ranking across interpretations or parse-trees. In serial

parsers, the top-ranked interpretation is chosen and the others discarded. In parallel parsers, some broader set of highly-ranked interpretations is maintained.

The two classes of parsing algorithms differ more substantially in modeling garden path effects. A serial parser models garden path effects by relying on some additional heuristics or innate parser structures which make the parser unable to build the correct interpretation. These structures might include the 3-constituent window of Marcus (1980), or the similar word-based window of the early Sausage Machine (Frazier and Fodor 1978).

In a parallel parser, by contrast, garden path effects are usually modeled as an effect of pruning. Memory or other constraints require the parser to prune some interpretations. Using the same ranking which models disambiguation preferences, the parser prunes some set of low-ranked interpretations. The garden-path effect is explained because the correct interpretation of an utterance is among these pruned interpretations.

In this section we argue that a parallel algorithm which ranks interpretations by their probability and prunes low-probability interpretations via beam-search can explain both preference effects and garden path effects. In addition, we show that the consequence of using probability as a metric is to choose interpretations which are the most coherent with previous expectations.

§4.1 first introduces relevant data on parsing preferences and garden-path sentences, and then §4.2 presents the **local coherence** model of disambiguation. §4.4 will compare our model with previous models of disambiguation and garden paths, arguing that the parallel probabilistic algorithm, besides accounting for the access results discussed in §3, accounts for a much broader variety of disambiguation results than previous models.

4.1 Psychological Results

The psychological results on disambiguation may be separated into two classes. First are studies, both on-line and off-line, on human disambiguation preferences. Examining which interpretations a subject prefers (of, say, an ambiguous prepositional phrase attachment) provides data to be modeled by a potential theory of disambiguation.

For example, Ford et al. (1982), in an off-line experiment, asked subjects to perform a forced choice between two meanings of an ambiguous utterance. They showed that in (21), subjects preferred to attach the prepositional phrase *on the beach* to the noun *dog*, while in (22), subjects preferred to attach the prepositional phrase to the verb *kept*.

(21) The women discussed the dogs on the beach.

a. The women discussed the dogs which were on the beach. (90%)

b. The women discussed them (the dogs) while on the beach. (10%)

(22) The women kept the dogs on the beach.

a. The women kept the dogs which were on the beach. (5%)

b. The women kept them (the dogs) on the beach. (95%)

Taraban and McClelland (1988) studied a number of preposition-attachment ambiguities, by measuring preferences when subjects had seen a sentence up to and including the preposition, but not including the prepositional-object head noun. In general, they found that subjects used both verbal and nominal expectations to try to attach the prepositional objects. They found for example in (23)–(25) that subjects preferred noun phrase attachments.

(23) The executive announced the reductions in the *budget / evening*.

(24) The philanthropist appreciated the story on his *generosity / deathbed*.

(25) The high-school senior stated his goals for the *future / principal*.

The second kind of psychological result deals with garden-path sentences. In garden-path sentences, readers are fooled by locally ambiguous structures into selecting an incorrect interpretation. We summarize below a number of cases of ambiguities that lead to the garden-path effect as well as some that do not. [6]

1. Main Clause–Relative Clause ambiguity is often resolved in favor of the main clause analysis.

   (26) # The horse raced past the barn fell. *(Bever 1970)*

   However, semantic context can reduce or eliminate the garden-path effect (Crain and Steedman 1985).

   (27) #The teachers taught by the Berlitz method passed the test.

   (28) ?The children taught by the Berlitz method passed the test.

2. Lexical category ambiguities sometimes cause garden path effects, sometimes do not.

   (29) # The complex houses married and single students and their families. (Hearst 1991)

   (30) The warehouse fires destroyed all the buildings. (Frazier and Rayner 1987)

   (31) The warehouse fires a dozen employees each year. (Frazier and Rayner 1987)

   (32) # The prime number few. *(Milne 1982)*

   (33) # The old man the boats.

   (34) # The grappling hooks on to the enemy ship. *(Milne 1982)*

3. Valence ambiguities sometimes cause garden path effects

   (35) # The landlord painted all the walls with cracks.

   (36) # Ross baked the cake in the freezer. *(from Hirst (1986))*

Any cognitive theory of disambiguation must explain why certain ambiguities cause processing difficulties and why others do not.

[6] We follow Gibson (1991) in marking garden-path sentences with a pound-sign (#)).

Indeed, recent psycholinguistic results have been used to argue that this distinction between disambiguation preferences and garden-path sentences is only an approximation, and that processing difficulty is better modeled as a continuum, with slight preferences at one end and garden-path-like pruning at the other (MacDonald et al. 1994; Tabossi et al. 1994). Since a probabilistic algorithm maps sentences onto the real numbers, it is well-suited to modeling this kind of continuum. In the absence of sufficient data on the details of the processing difficulty continuum, however, the rest of this section will simply show that our coherence-based algorithm models the parse preference effects, and that the same preference ranking together with a pruning algorithm can be used to predict the garden-path effect for the garden-path data.

4.2 The Model

§3 showed that a parallel algorithm which uses probabilities to rank constructions could be used to model lexical and constructional access. This section shows that the same probabilistic ranking can be used to account for disambiguation preferences. Each interpretation will be ranked by its probability, computed from both constituent and valence probabilities. We show this ranking is able to account for the psychological data on disambiguation preferences.

A disambiguation model must also explain the processing difficulties that garden-path sentences cause human parsers. The serial parsing models that were traditional in psycholinguistics model garden path effects by relying on some particular heuristics or innate parser structures which make the parser unable to build the correct interpretation. For example, because of its 3-constituent window, Parsifal (Marcus 1980) is unable to simultaneously view the initial noun phrase 'the horse' and the final word 'fell' in (26). Since the parser is required to build structure before 'the horse' moves out of its window, and since there is no evidence yet for the reduced-relative parse, Parsifal will build the main-verb parse. Thus a hardwired memory constraint is used to force the parser to drop the correct parse from consideration.

In a parallel parsing model, however, there is a natural alternative to these specific heuristics or parser structures, proposed by Kurtzman (1985), Norvig (1988), and Gibson (1991). In their models, garden-path sentences are explained by the the same ranking that is used to model disambiguation preferences. The parser ranks each hypothesis, and then relies on some sort of constraint (such as limited-memory) to force low-ranked hypotheses to be pruned. The garden-path effect is explained because the correct interpretation of an utterance is among these pruned hypotheses.

In our model, the parser ranks each parse by its probability, and uses beam-search to limit the search-space. As Gibson (1991) proposes, we can find an empirically acceptable beam width and also test our model by considering the probabilities of each interpretation of ambiguous sentences. A sentence where there is evidence for pruning (such as the garden-path effect) sets an upper bound on the beam width. A sentence where there is no evidence for pruning (both interpretations are acceptable) sets a lower bound on the beam width. We show that a beam-width ratio of anywhere between 3.8:1 and 5.6:1 between the best and the pruned hypothesis is sufficient to account for a number of garden-path examples. Table 1 summarizes some of the garden

path/embedded anomaly and non-garden path examples which we will work through in the rest of this section In the first three examples (explained in further detail below; †indicates a local embedded anomaly) the ratio between the probabilities of the best and pruned hypothesis is 5.6 or greater. In the last two examples, where the local ambiguity does not cause a garden path, the ratio is 3.8 or less.

#*The complex houses* married and single students and their families	267:1
†*The sheriff wasn't sure which rock the cowboy raced...* (desperately past)	12:1
†*The district attorney found out which church the reporter asked...* (anxiously about)	5.6:1
The warehouse fires destroyed all the buildings	3.8:1
The bird found in the room died.	3.7:1

Table 1: Choosing a Beam-Width: Probability Ratios

We turn now to the actual computation. §2 summarized the two kinds of probabilities that augment constructions, constituent probabilities and valence probabilities. The constituent probability is the standard SCFG rule-probability, and as we discussed an SCFG assumes that the probability of each rule or construction is independent, and hence that the probability of a parse tree can be computed by multiplying the probabilities of the individual rules. We also make the simplifying assumption that valence probabilities are independent of constituent probabilities and of each other. Thus the likelihood of a sentence given an interpretation is the product of each constituent and valence probability in its derivation. Figure 7 shows the probabilities associated with a small noun phrase beginning a sentence.

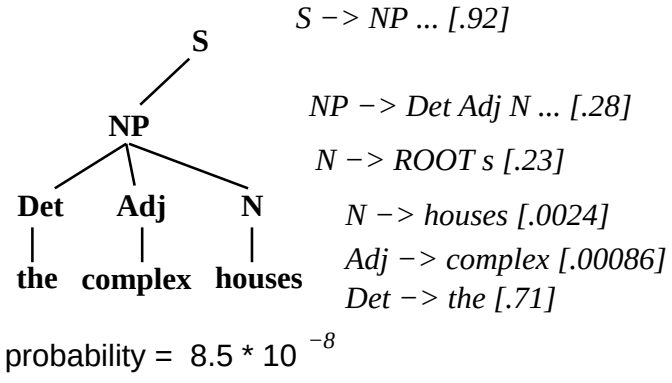


Figure 7: Computing the probability of a derivation

4.3 Modeling Preference Effects

Lexical Valence Preference

Consider the Ford et al. (1982) results on attachment preferences for *keep* and *discuss* summarized above. Ford et al. explained their results by proposing that keep and discuss each had two possible subcategorization frames; one with a single <NP> complement, one with <NP PP>. They proposed that the verbs differed, however, in

their preferences for these valence structures; *keep* prefers <NP PP>, while *discuss* prefers <NP>.

Our interpretive model is conveniently set up to test Ford et al.'s intuitions, since both constructions and valence structures are annotated with conditional probabilities. We assume that *keep* and *discuss* each have two valence frames. For *discuss*, we assume for simplicity that the two frames are defined strictly syntactically as <NP> and <NP PP>. For *keep* the frames are somewhat more complex to define. While the optional second argument of *keep* can be a PP (keep the dogs *on the beach*), Table 2 shows that it can also be an AP, a VP, or a particle.

AP	keep the prices reasonable
VP	keep his foes guessing
VP	keep their eyes peeled
PRT	keep the people in
PP	keep a number of surprises under our hats
PP	keep his nerves from jangling

Table 2: Complements of *keep*

Most theories model this class of complements syntactically, as *secondary predicates*, represented as *[pred +]*, although it is also possible to capture them thematically as RESULTs. Assuming, again only for simplicity, the *[pred +]* representation, *keep* has the thematic grid options <NP> and <NP XP[pred +]>.

We used the Penn Treebank to compute probabilities for each valence structure. The results are shown in Table 3.

Table 3: Valence Probabilities computed from the Penn Treebank

discuss	<NP PP>	*.24*
	<NP>	*.76*
keep	<NP XP[pred +]>	*.81*
	<NP>	*.19*

If we only consider these probabilities, we arrive at a result consistent with the Ford et al. (1982) data. Since *discuss* prefers a lone NP complement, the PP will attach to the noun phrase, while *keep* prefers to attach the PP as an argument of the verb.

However the different attachments of the preposition phrases also lead to different phrase structures. Figure 8 shows the different phrase structures for each interpretation. The interpretation in which *on the beach* is a complement of 'keep', is shown above. This interpretation includes the VP-level rule with a slot for oblique complements: [7]

(37) $[.15]\ VP \rightarrow V\ NP\ XP$

The probability of this interpretation includes this probability, .15, as well as the valence probability that *keep* fills its RESULT argument, or .81. All the other probabilities

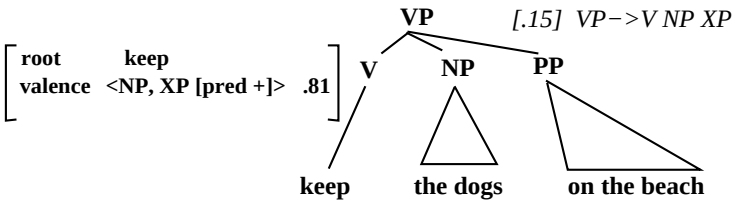
[7] We assume here a GPSG-like treatment of VP rules; generalizations across different VP types are captured by an abstract VP in the type hierarchy.

involved in the interpretation are shared with the other derivation, and so we need not consider them when comparing the two. The second interpretation has a slightly more complicated phrase structure at the NP level, and so there are two rules in this parse that are not in the other:

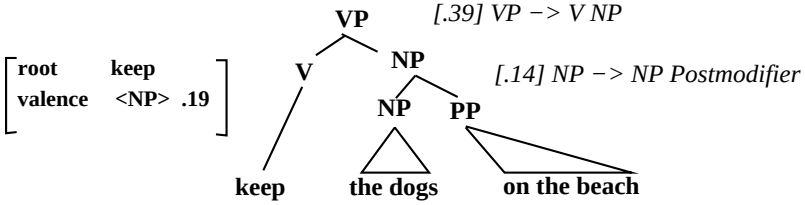
(38) $[.39]$ *VP* $\rightarrow$ *V NP*

(39) $[.14]$ *NP* $\rightarrow$ *NP XP*

Note that the probability of the simple transitive VP is higher than the probability of the multiple-complement VP above. This is offset, however, by the lower probability assigned to the unfilled RESULT valence slot, and in addition by the lowered probability due to the extra NP rule. Thus the probability of the verb attachment is .12/.01= 12 times the probability of the noun attachment, roughly modeling the Ford et al. (1982) results, in which the verb attachment was preferred 19 times over the noun attachment (95%).



(a) .15 * .81 = .12 (prefered)



(b) .19 * .39 * .14 = .01 (disprefered)

Figure 8: Annotated Parse Trees for Two Interpretations of *keep the dogs on the beach*

Figure 9 shows the two interpretations of *discuss the dogs on the beach*. Although the syntactic structures are the same, the different valence probabilities of *discuss* and *keep* will cause the model to prefer the opposite interpretation. Although the prediction is correct, the model's 53% preference for the NP attachment does not quantitatively match the 90% preferences in the Ford et al. (1982) results. One possible explanation for this is the forced-choice nature of the Ford et al. experiment. Requiring the subjects to choose between the two interpretations may have caused the experiment to overestimate the probability of the most-preferred interpretation.

Augmenting valence slots with probabilities also allows us to explain with no further assumptions the stronger preferences for obligatory over optional arguments. Obligatory arguments will have unity or near-unity probabilities, while optional arguments will have lower probabilities. Britt (1991) showed, for example, that an obligatory

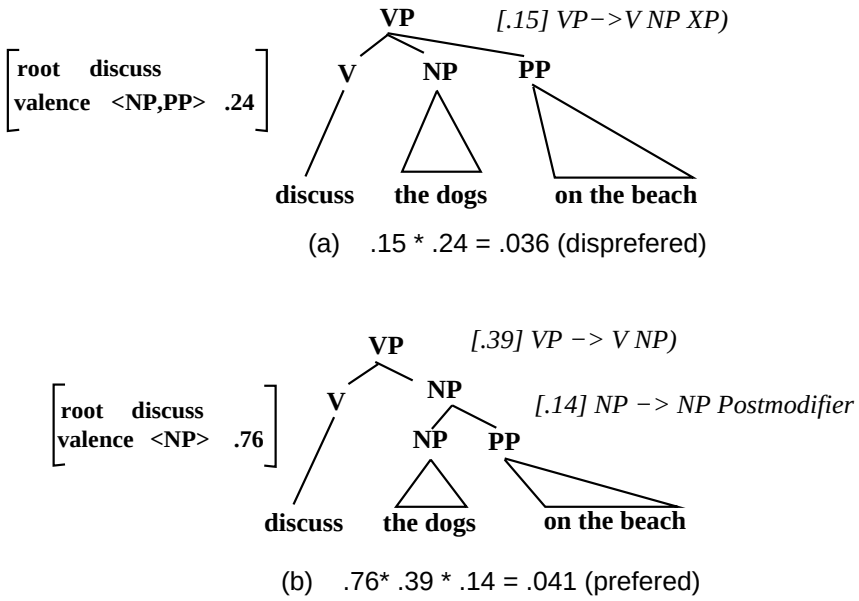


Figure 9: Annotated Parse Trees for Two Interpretations of *discuss the dogs on the beach*

argument slot placed a stronger expectation to be filled than an optional one. In an on-line experiment, Britt showed that when choosing between noun and verb attachment of ambiguous preposition phrases in two-referent contexts, subjects preferred to attach the PP to the verb if the verb obligatorily subcategorized for the PP. If the verb only had an optional subcategorization for the PP, there was a preference for noun attachment instead.

Viewing the probabilistic model in a more general way, we can explain the general preference for arguments over adjuncts as resulting from a combination of valence probabilities and constituent probabilities. The arguments will have a non-zero valence probability, while adjuncts do not fill a valence slot in a predicate, and so not attaching an adjunct incurs no probabilistic penalty.[8] This is true for both nominal and verbal adjuncts. In addition, when choosing between verbal arguments and nominal adjuncts, since nominal adjuncts will require an extra stochastic context-free rule to adjoin with, they will have a lower probability than a simpler structure that has no such rule. This is because due to the independence assumption inherent to context-free grammars, the probabilities of each rule are multiplied together. Thus a more complex structure will have more probabilities to multiply, and since each probability is less than 1, a lower total probability. Note that the fact that a more complex structure is likely to have a lower probability than a simpler structure may explain the success of heuristics like Minimal Attachment, which preferred attachments producing the simplest syntactic structure. However, a simpler structure will not necessarily have a higher probability, since it is not just the cardinality of the probabilities but their values which matter.

In order to explain this sort of preference, Ford et al. (1982) and Abney (1989) proposed a heuristic that preferred attachments to verbs over attachments to nouns. Relying

[8]This is especially true if we assume the common model in which complements are the arguments of verbs but adjuncts act as predicates in their own right.

on valence probabilities eliminates the need for this occasionally faulty heuristic.

4.3.1 Modeling Garden Path Effects

In our model, as in most explanations for the garden-path effect, the effect arises when the processor chooses an interpretation which fulfills local linguistic expectations over one which is globally plausible, pruning the more plausible interpretation. Testing this model requires showing that the theory predicts pruning in a case of ambiguity whenever the garden-path effect can be shown to occur. If the beam width is set too wide, the theory will mislabel a garden path sentence as merely a less preferred interpretation. Conversely, if the window is too narrow, the theory will mislabel parsable sentences as garden-paths.

In many ways we view our model as an extension of the parallel model of Gibson (1991). Gibson first showed that a parallel (although non-probabilistic) model could account for the wide range of syntactic garden path effects, and proposed that probabilistic garden-path effects might also be modeled with a parallel algorithm. In the rest of this section, we work through some examples of garden paths, showing that the coherence model can not only account for garden paths that seem obviously probabilistic in nature (such as (29) and (33) above), but also more traditional garden path sentences such as (26). We consider these garden paths in three classes: those caused by constituent probabilities, those caused by valence probabilities, and those caused by combinations of the two.

We will show that a beam-search in which interpretations are pruned if they have less than about 1/5 the probability of the best interpretations can account for all of the data we consider.

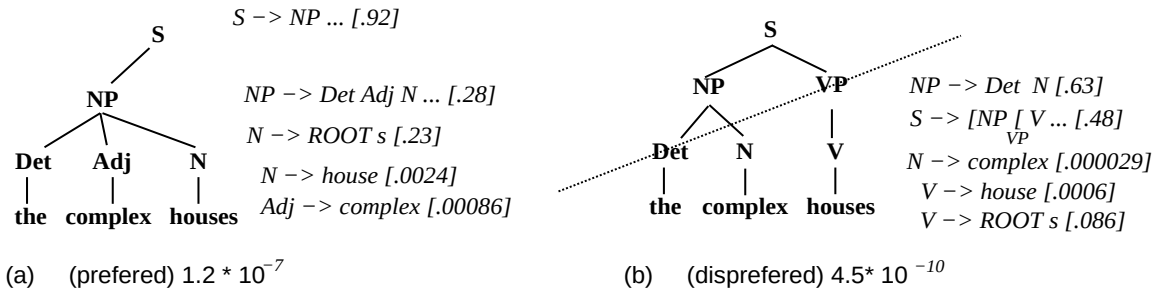
Garden Paths Caused by Construction Probabilities

For simplicity, we begin by considering garden paths caused by simple phrase-structure probabilities. Consider the garden path sentence (40). In the intended interpretation of the sentence, 'complex' is a noun, and 'houses' is a verb; thus the students are housed by the complex. However, most readers initially interpret 'the complex houses' as a noun phrase, and are confused by the lack of a verb.

(40) # The complex houses married and single students and their families. (Hearst 1991)

Figure 10 shows the two interpretations of the prefix *the complex houses*. For convenience (and in order to make the model maximally general) we have shown the phrase structures in a simplified minimal-assumption framework.

In the initially preferred interpretation in (a), 'houses' is the head noun and 'complex' is an adjective. In the dispreferred interpretation in (b), 'houses' is a verb and 'complex' the head noun. The difference in probability between the two interpretations is due mainly to the different lexical category probabilities. (41)–(46) show conditional probabilities generated from the Brown Corpus. Notice that 'houses' is more likely to be a noun than a verb, while 'complex' is more likely to be an adjective than a noun.



(a) (prefered) $1.2 * 10^{-7}$ (b) (disprefered) $4.5* 10^{-10}$

Figure 10: Annotated Parse Trees for Two Interpretations of *The complex houses*

In addition, we have shown the probabilities of the plural and 3rd-person-singular interpretations of the ambiguous 's' morpheme.

(41) $[.0024]\ N \rightarrow$ *house*

(42) $[.0006]\ V \rightarrow$ *house*

(43) $[.00086]\ Adj \rightarrow$ *complex*

(44) $[.000029]\ N \rightarrow$ *complex*

(45) $[.23]\ N \rightarrow$ ROOT *s*

(46) $[.086]\ V \rightarrow$ ROOT *s*

The non-lexical nodes of the parse tree seem irrelevant here because the rest of the trees differ in probability only by about 20%. Computing the complete probabilities of the two interpretations correctly predicts the garden path effect; first, the probabilities correctly prefer the (a) interpretation, and second this reading is 267 times more probable than the dispreferred reading. Since this is much greater than the beam width, interpretation (b) will be pruned.

By contrast, (47) shows an example due to Frazier and Rayner (1987) where a similarly ambiguous construction does not cause the garden path effect.

(47) a. The warehouse fires destroyed all the buildings.

b. The warehouse fires a dozen employees each year.

Note that the difference between (40) and (47) cannot be due to valence differences or non-probabilistic structural differences, since the two examples have very similar valence and syntactic structure. The difference between the two examples is in the lexical probabilities.

The difference in probability of lexical category for *fire* is shown in (48)–(49). The two interpretations in Figure 11 have probabilities of $4.2 * 10^{-5}$ and $1.1 * 10^{-5}$; their ratio of 3.8/1 is less than the beam-width, and hence neither interpretation will be pruned.

(48) $[.00072]\ N \rightarrow$ *fire*

(49) $[.00042]\ V \rightarrow$ *fire*

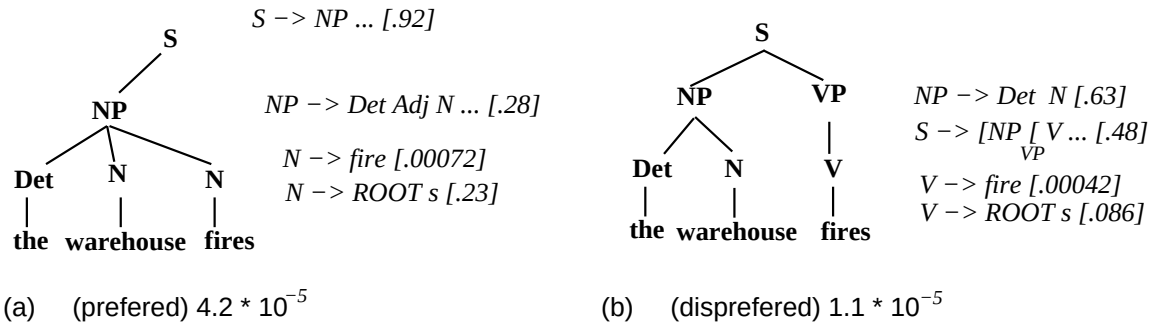


Figure 11: Annotated Parse Trees for Two Interpretations of *The warehouse fires*

Garden Paths Caused by Valence Probabilities

For an example of the garden path effect caused by valence probabilities, we turn to an experiment which was originally designed to test Fodor's (1978) Lexical Expectation Model of gap-finding in a serial parser. Fodor was attempting to model how the processor decided which gaps to fill with a given fronted element. For example, the fronted NP *the book* in (50a) could turn out to be the object of *want* or the object of some embedded verb like *buy*.

(50) a. I saw the book you wanted...
b. *Book is object of want*: I saw the book you wanted.
c. *Book is object of deeper verb*: I saw the book you wanted me to buy.

Fodor proposed that the human sentence processor posits gaps only following verbs which are *more likely* to be transitive. That is, given a fronted argument, when the parser reaches a verb it should immediately attach the argument as the direct object of the verb if the verb is more likely to be transitive, otherwise it should wait. Assuming that *ask* is preferentially transitive and *race* preferentially intransitive, the parser would attach 'the man' as a direct object in *the man you asked*, but wait in *the book you raced* (assuming rather that 'the book' should be attached as the object of some lower predicate as in *the book you raced past*). Tanenhaus et al. (1985) tested this idea by creating test sentences with fronted arguments which make sense if the whole sentence is seen, but which are semantically anomalous direct objects of the verb. Thus if the parser tries to fit them into the verb right away, the reader should have a minor processing breakdown. If Fodor's model is correct, the parser should suffer an anomaly when trying to fit a semantically implausible filler into preferentially transitive verbs like *ask*, but should not break down on preferentially intransitive verbs like *race*. The experiment used sentences like 51–52, testing for an anomaly effect at the verb.

(51) a. The district attorney found out which witness/church the reporter asked —— about the meeting. (*early gap*)
b. The district attorney found out which witness/church the reporter asked anxiously about ——. (*late gap*)

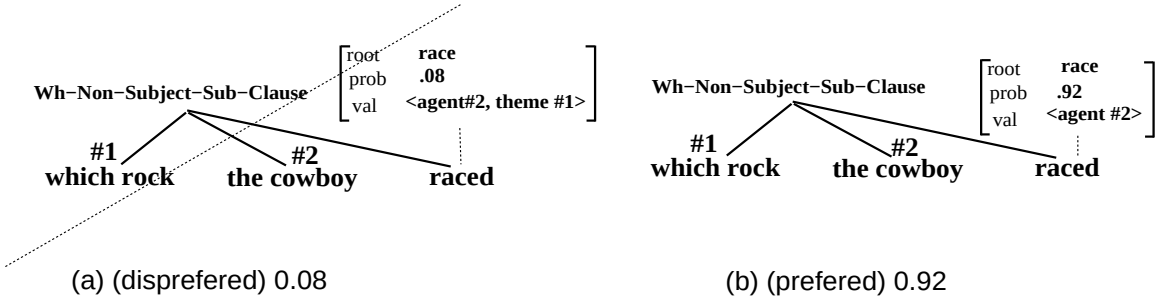
(52) a. The sheriff wasn't sure which horse/rock the cowboy raced —— down the hill. (*early gap*)

b. The sheriff wasn't sure which horse/rock the cowboy raced desperately past ——. (*late gap*)

The results of the experiment supported Fodor's model; for transitive-preference verbs, there was an anomaly effect at the verb, while for intransitive-preference verbs, there was no anomaly effect at the verb. Tanenhaus et al.'s (1985) explanation was based on a serial architecture; for verbs which are preferably transitive, the parser hypothesizes a gap; for verbs which are preferably intransitive, the parser does not hypothesize a gap.

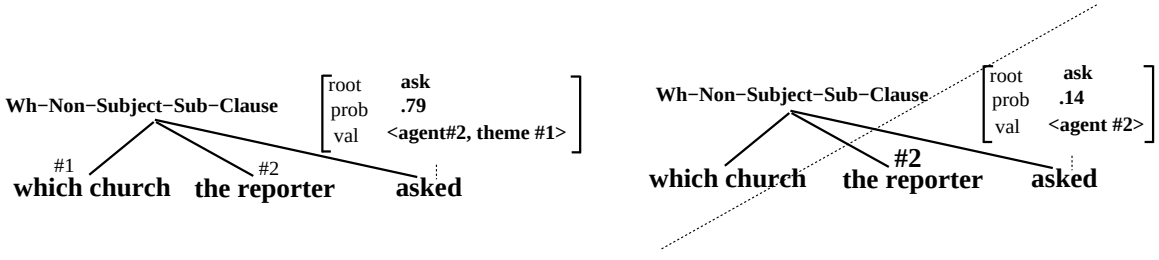
We argue that rather than requiring a special gap-hypothesizing mechanism, that these results can be accounted for with our parallel architecture without proposing any special mechanism. First, as with the other valence ambiguities we have seen, the parser will activate both possible valence interpretations for each verb, one transitive and one intransitive. However, in each case the low-probability interpretation will be pruned. Because the low-probability intransitive interpretation of *ask* will be pruned, the parser will attempt to bind *church* as the direct object of *ask*, causing an anomaly. Because the low-probability transitive interpretation of *race* will be pruned, the parser will not attempt to bind *rock* as the direct object of *race*, avoiding the anomaly.

Figure 12 and Figure 13 show the disambiguations for *race* and *ask* respectively. The valence probabilities were determined from Connine et al. (1984). For *race* the ratio of the two interpretations is .92/.08 = 12/1. For *ask* the ratio is .79/.14 = 5.6/1. In both cases this is sufficient to cause pruning of the dispreferred interpretation.



(a) (disprefered) 0.08 (b) (prefered) 0.92

Figure 12: Access both thematic grids in parallel, transitive is pruned



(a) (prefered) 0.79 (b) (disprefered) 0.14

Figure 13: Access both thematic grids in parallel, intransitive is pruned

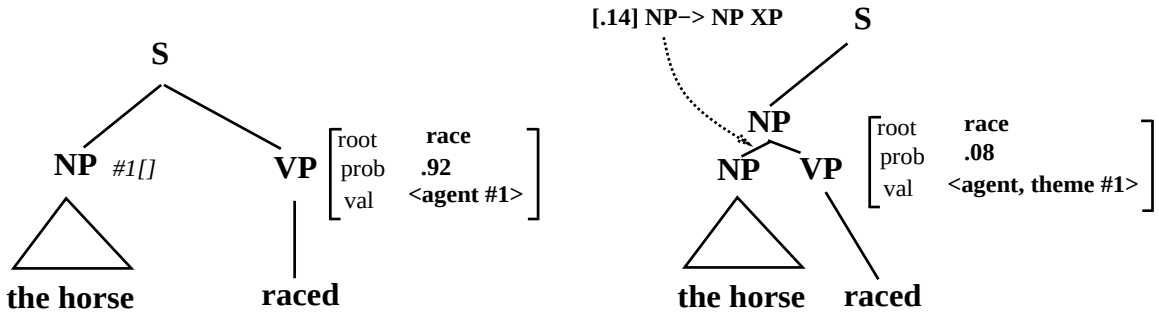
More on Horse-racing: Combinations of valence and construction probabilities

We have now seen examples of garden-path sentences caused by construction probability differences, and by valence probability differences. In this section we argue that Bever's (1970) familiar garden-path sentence in (53) is caused by a combination of both valence and construction probabilities.

(53) #The horse raced past the barn fell.

Here the phrase *raced past the barn* is ambiguous between a reduced relative clause VP and a main-clause VP. Clearly the garden path effect arises because the main verb reading of *race* is somehow preferred by the parser, while the reduced-relative clause turns out to the be correct one.

We show first that the preference for the main-verb reading arises from our probabilistic model, and second that the difference between the two interpretations is sufficient to cause the reduced-relative interpretation to be pruned.



(a) main verb interpretation

(b) reduced relative interpretation

Figure 14: Pruning the reduced relative interpretation

Figure 14 shows the two relevant interpretations of the sentence prefix *'the horse raced'*. Figure 14(a) shows the main verb interpretation. Note that the intransitive valence of *race* has a probability of .92.[9]

The reduced relative interpretation requires the transitive valence of *race*, since it requires the first NP to be bound as an object of the verb. In addition, this interpretation has a more complex syntactic structure; note that in addition to the syntactic rules used in the trees in (a), the tree in (b) has an additional rule:

(54) [.14] NP → NP XP

The combination of the extra (and low-probability) rule and the lower valence structure makes this interpretation 82 times less probable than the best-ranked main-verb interpretation. Since this places it outside the 5x beam width, the reduced-relative interpretation will be pruned.

Early attempts to explain the garden-path effect of this sentence relied solely on its syntactic structure. Bever's (1970) original proposal relied on a heuristic to prefer

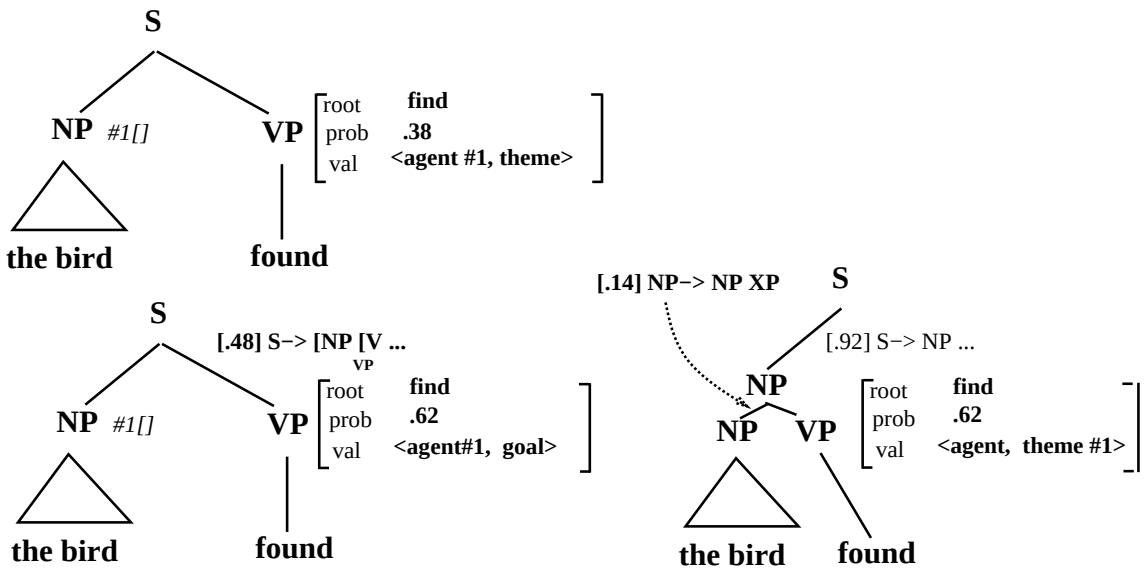
[9]These numbers are from Connine et al. (1984).

main verbs over reduced verbs, Marcus's (1980) relied on the fact that a 3-constituent window would not include the final verb, and heuristics like Minimal Attachment (Frazier and Fodor 1978) rely specifically on the parse tree.

However Pritchett (1988) recently showed that solely constituent-based solutions to the problem are insufficient, as sentences like (55) do not cause the garden-path effect, despite having exactly the same syntactic structure as (53).

(55) The bird found in the room died.

The difference between (55) and (53) is the valence structure of the verbs. Where *race* is preferably intransitive, *found* is preferably transitive. Figure 15 shows the three interpretations of the prefix *'the bird found'*.



(a) two main verb interpretations

(b) one reduced relative interpretation

Figure 15: Not every reduced relative causes Garden Path effects

Again, the two main-verb interpretations are too close to each other to cause pruning. Significantly, however, the difference between the main and reduced-relative interpretations is much less than it was for *race*, the probability ratio being (.62*.48)/(.62*.14*.92) or 3.7/1. This ratio is less than the beam width, and so the reduced relative interpretation will not be pruned.

4.4 Recent Theories of Disambiguation

If every theory of disambiguation relies at its core on some metric for comparing interpretations, theories may be classified by the kind of metrics they employ. Many models use *global metrics*, in which each structure is ranked by a single global criterion, such as our ranking based on conditional probability, or such as a preference for the most semantically plausible interpretation. Such global metrics are particularly

commonly used by the interactive or constraint-based models of sentence processing such as the model presented here as well as others like MacDonald (1993), McClelland et al. (1989), Spivey-Knowlton and Sedivy (1995), and Trueswell and Tanenhaus (1994). In contrast, *local structural* models choose between interpretations by using local, domain-specific heuristics based on simple structural patterns, such as choosing local attachments, or choosing syntactically simpler structures. These local metrics are commonly employed in the two-stage parsing models, in which a modular syntactic processor makes initial disambiguation decisions based only on local syntactic information, without access to lexical and semantic knowledge. These decisions may then be reanalyzed by a second, knowledge-rich stage of processing.

This section first surveys two classes of global heuristics, those based on semantic plausibility and those based on syntactic or textual coherence. We will argue that the semantic plausibility model of Crain and Steedman's (1985) and others cannot account for the full range of experimental results on garden-path effects, while Pritchett's (1993) coherence-based model is insufficiently general to account for lexical access effects and frequency effects. Finally, we compare our our probabilistic algorithm to other parallel algorithms (Gibson 1991; Spivey-Knowlton 1994). We next consider the local heuristics; as §4.3 suggested, a preference for minimal structures, all else being equal, falls out of the independence assumptions inherent to context-free grammars. We focus in this section, therefore, on the locality heuristics, arguing that, although the human parsing clearly displays some effects of locality, no previously proposed locality heuristic is sufficient to account for the range of data.

4.5 Global Metrics

The most well-known global heuristic for modeling human parsing preferences and garden-path data, and one that has perhaps the most intuitive appeal, is simply to choose the most plausible interpretation. This model was expressed most familiarly by Crain and Steedman (1985:330):

> *The Principle of A Priori Plausibility*. If a reading is more plausible in terms either of general knowledge about the world, or of specific knowledge about the universe of discourse, then, other things being equal, it will be favored over one that is not.

A number of researchers have proposed models along these lines, including Kurtzman (1985), Altmann and Steedman (1988), and Charniak and Goldman (1988). One disadvantage of this model is that in most cases it is hard to see how to make it operational. For example, what is it about having a discussion about dogs which are on the beach that makes it more plausible than having a discussion on the beach about dogs? The model also has serious difficulties in accounting for garden-path sentences. For example, Crain and Steedman (1985) argue that the preference for the main clause interpretation in main-clause/reduced-relative ambiguities is a corollary of plausibility. This because the reduced relative clause reading requires more presuppositions than the main clause reading (in particular the presupposition in 'the horse raced past the barn fell' that there are multiple horses). However, this model incorrectly predicts the

garden path effect for (56):

(56) The bird found in the room died.

The problem is that the plausibility model does not take into account verbal valence preferences. Finally, the model cannot account for the frequency effects summarized above. This is not to deny the role of plausibility in parsing data; as §5 suggests, one advantage of our probabilistic model is that in principles it allows the inclusion of plausibility information along with any other sources of evidence in ranking interpretations.

Most other global or global-like models use something like coherence as the metric for ranking interpretations. The roots of coherence as a processing heuristic lie in Wilks' (1975b, 1975a) rediscovery of the *Joos Law* (Joos 1972), which argued for choosing a meaning which was most redundant and hence most coherent with the context (see also Hill (1970) and Joos (1958)). The idea was then taken up by the coherence-based marker-passing algorithms (Hirst and Charniak 1982; Norvig 1987; Hirst 1986). More recently, Ford et al. (1982) added the insight that verbal valence was the key factor in modeling parsing preferences and garden-path sentences, incorporating a preference for argument attachments over adjunct attachments, and also proposed preference augmentations to valence.

The modern models which account for the most data are Gibson (1991) and Pritchett (1993). Both of these models, like our probabilistic approach, emphasize coherence as a parsing heuristic. Pritchett's model is based on a serial parser which often reanalyzes its previously-produced parse, but which garden-paths in certain kinds of reanalyses:

> **Theta-Reanalysis Constraint**: Syntactic reanalysis that reinterprets a θ-marked constituent as outside of current θ-domain is costly.

When faced with an ambiguity, Pritchett's parser chooses the most θ-coherent parse; i.e. the one in which the most θ-grid slots are filled, and the most arguments have θ-roles. The model accounts for 'the horse raced past the barn fell', because *the horse* is originally parsed as the subject of *race,* and hence within *race*'s θdomain. When the parser sees *fell,* it will need to reanalyze *the horse* as within the θdomain of *fell*, causing the garden-path effect.

Although Pritchett's model accounts for a wide variety of syntactic garden-path effects, it is unclear how it would be extended naturally to account for the probabilistic effects in (40), or the semantic or context effects in (27)–(28). In addition, of course, Pritchett's model requires disambiguation to be completely independent of access and its frequency effects, despite the evidence that lexical disambiguation at least is strongly affected by access frequencies.

Gibson (1991) proposed a parallel parsing model in which each parse of an ambiguous input was ranked by its *processing* load. Interpretation proceeded by beam search; parses were pruned if their load was significantly higher than the best interpretation. Load was defined with three separate load parameters:

1. Property of Thematic Reception: parses have processing load for each argument lacking a θ-role

2. Lexical Requirement: parses have processing loads for each θ-grid slot lacking a filler.

3. Thematic Transmission: parses have processing load if a θ-role is passed through a chain (i.e. wh-movement or comp)

Like Gibson's model, our probabilistic model will still disprefer interpretations if they have unfilled slots or unparsed arguments. In a generalization over Gibson's model, however, our model can distinguish between two interpretations both of which have valence slots filled, but with different probabilities. In addition, our model can explain why the parser will sometimes prefer not to fill a slot with an argument, as in (57); as we argued in Figure 12, in this sentence because 'race' is preferably intransitive, the NP 'which rock' does not fill the direct object slot in 'race'.

(57) The sheriff wasn't sure which rock the cowboy raced . . .

Our explanation of reduced-relative/main-clause ambiguities is also different from Gibson's, although both models rely on the valence of the main verb to account for *race/found* distinctions. Where Gibson appeals to the Property of Thematic Transmission to account for part of the difficulty of 'The horse raced past the barn fell', our model appeals to the low prior probability of the reduced relative clauses. The advantage of a probabilistic account, again, is that it explains the garden-path examples based on frequency with the same mechanism as the structural garden-paths. Although Gibson points out that his parallel model could easily be extended to deal with frequency effects, it would presumably require adding another property or set of properties; our theory accounts for the data with a single model.

In our own model, as we've described it so far, as well as in the Gibson and Pritchett architectures, processing difficulty is a binary effect of parse pruning; sentences are either difficult to process or not. Because of this, the model is unable to account for processing difficulty variation within garden-path sentences (MacDonald et al. 1994; Tabossi et al. 1994). The probabilistic competition model of Spivey-Knowlton (1994) and Spivey-Knowlton and Tanenhaus (1995) attempts to address this lack by modeling processing difficulty as a continuum, caused by the time-course of competition between syntactic alternatives. Spivey-Knowlton and Tanenhaus (1995) combine information from discourse (in terms of the discourse context probability of a reduced relative), from verb frequency information (the probability of a verb appearing in the simple past tense form versus the past participle form), and from parafoveal information (the presence of "by" after the verb). They estimate probabilities for each of these information sources, and then combine them (via recurrent feedback and normalization) on a set of sentences with reduced-relative/main-clause ambiguities. They show that the durations of competition between the syntactic alternatives predicts the processing difficulty in first pass reading times at the verb.

Since our probabilistic metric already maps each possible interpretation of an ambiguous sentence onto a continuous space of probabilities, we hope to investigate whether our algorithm could incorporate the insights of competition-based approaches like that of Spivey-Knowlton and Tanenhaus (1995) or Stevenson (1993) in order to model the continuum of processing difficulty.

4.6 Structural Heuristics

Structural heuristics were originally proposed by Kimball (1973) to explain the ungrammaticality of certain sentences, or the inability of subjects to get some readings of an ambiguous sentence. Since then a number of heuristics have been proposed, falling into two classes: heuristics to build the syntactically simplest structure (*minimality heuristics*) and those to combine nearby structures (*locality heuristics*).

However, many authors (Kurtzman 1985; Norvig 1988; Gibson 1991; Schubert 1986; Osterhout and Swinney 1989) have noted that it is quite easy to choose particular lexical items or particular contexts which reverse any of the heuristics, particularly the minimality heuristics. Thus more recent models incorporate local heuristics in a more sophisticated way; as the first stage in a two-stage parser. This first stage is an encapsulated module which makes a very early, local decision about structure. This decision is reviewed, and possibly reanalyzed, by a second stage which has access to richer knowledge sources. A number of papers have argued that only minimality and locality are relevant to this first-stage process, and that garden-path data and grammaticality judgements have no access to this early stage of processing (Clifton 1993; Clifton et al. 1991; Mitchell 1989; Ferreira and Clifton 1986). Trueswell and Tanenhaus (1994), in contrast, have argued that results such as those of Ferreira and Clifton (1986) can be reinterpreted to argue for a more constraint-based view of processing.

A resolution of these experimental differences awaits further experimental work; however, in the rest of this section we argue that formulating any single locality heuristic which explains even the available experimental data will be quite difficult. One major weakness of minimality heuristics, for example, is that they are dependent on quite particular assumptions about the grammar. As Norvig (1988) points out, for example, a syntactic simplicity heuristic like Minimal Attachment makes no sense in a categorial grammar, in which every derivation has the same number of nodes. In addition, since the local heuristics only account for garden-paths and preferences due to syntactic effects, a two-stage theory would require other, presumably unrelated mechanisms to account for garden paths due to frequency effects or to non-local and semantic expectations. An interactive or constraint-based approach to disambiguation accounts for both syntactic and non-syntactic effects with a single mechanism.

Another problem with the use of local heuristics and the serial disambiguation strategies that usually accompany them is the well-accepted evidence that disambiguation is not always immediate. Frazier and Rayner (1987), for example, showed that the parser was able to delay disambiguation in (47) above (repeated in (58) below), avoiding the garden path.

(58) a. The warehouse fires destroyed all the buildings.

b. The warehouse fires a dozen employees each year.

The only way to model these effects in a serial parser is to add a complex delay function, for which there is no independent motivation. In a beam-search parallel parser, on the other hand, delay is simply caused by the fact that the two interpretations in (58) have probabilities within the beam. The existence of lexical ambiguities like (58) which do

not cause the garden-path effect also argues against Ford et al.'s (1982) proposal that it is lexical category reanalysis which explains the difficulty of recovering from garden paths.

In the remainder of this section we present a paradox for locality heuristics. We show that there are two styles of locality heuristic, and that experimental results and grammaticality judgements which have been taken as evidence for one kind of locality heuristic cannot be modeled by the other. We are not arguing that locality plays no role in language processing; it clearly does. Besides the experimental results discussed above, for example, Gibson and Pearlmutter (1994) found evidence for locality heuristics in a statistical analysis of preposition-phrase attachment in the Brown corpus. However, we argue that although each version of locality accounts for some data, there are principled problems with generalizing any of the locality models to handle all the data. Thus the intuition that locality plays some role in processing is very difficult to make operational.

Locality principles described in the literature include **Right Association** (Kimball 1973), **Local Association** (Frazier and Fodor 1978), **Late Closure** (Frazier 1978), **Final Arguments** (Ford et al. 1982), the **Graded Distance Effect** (Schubert 1984) & 1986, **Rule B** (Wilks et al. 1985), **Attach Low and Parallel** (Hobbs and Bear 1990), and **Recency Preference** (Gibson 1991),(1993). These models of locality fall into two classes. In the **windowing** models, locality effects are explained by appealing to a limited buffer or window on the input sentence, generally assumed to model some sort of limited memory. Since the parser is limited to working within this window, constituents will necessarily be attached locally if they can be. The window may be limited by the number of words (Frazier and Fodor 1978; Magerman and Marcus 1991) or the number of constituents (Marcus 1980).

In the **iterative** or **right-to-left** models, when attempting to attach a constituent, the parser moves right to left until it finds the first appropriate head to attach the constituent to. Unlike the windowing models, in these models locality does not fall out of the structure of the parser, and must be stipulated. Iterative models differ in how they define the notion of *appropriate head* – this may be semantically appropriate (Wilks et al. 1985; Hobbs and Bear 1990; Whittemore et al. 1990) or syntactically appropriate (Gibson 1991).

For example, the *Local Association* principle of Frazier and Fodor (1978), is a windowing model. It establishes a fixed-length buffer which can hold five or six words, and predicts that locality effects can be explained by the limited view imposed by this buffer. While windowing models have the advantage of *explaining* locality rather than stipulating it, it has proven to be very difficult to satisfactorily define a window size, as we will see below.

The *Recency Preference* principle of Gibson (1991) is an iterative or right-to-left model. In this model, whenever there is more than one possible attachment point for an adverbial, all but the most recent attachment point are removed from consideration. In his case, adverbials can be attached to verbs or sentences, so the Recency Preference principle requires that an adverb cannot "skip" a local verb and attach to a more distant one.

Another model, again an iterative one, assumes what might be called the *Most*

Recent Semantically Compatible Attachment principle, first proposed by Wilks et al. (1985) (as the *Rule B* algorithm of the CASSEX program), and used also by Hobbs and Bear (1990) and Whittemore et al. (1990). These models choose among attachments by attempting to attach an adverbial to each possible head, starting with the most recent, and moving further left, and selecting the first one that fits semantically.

We discuss the performance of these models on three classes of data: garden-path sentences, restrictive relative clause attachment, and verb-particle attachment.

A number of parsing models have attempted to use windowing locality to explain the garden-path effect (although this has not been proposed with iterative models). Consider (59):

(59) #The horse raced past the barn fell

All windowing models predict the difficulty of this sentence by arranging that only *The horse raced past the barn* is in the window. In the Marcus (1980) parser, the three constituent buffer is filled by *the horse, raced*, and *past the barn*; in the Local Association model the buffer is exactly six words long.

The problem with the windowing models, as has been noted by Norvig (1988), Gibson (1991), Schubert (1986), and Pritchett (1993), is that for any given window size, it has been shown possible either to construct a parsable sentence which does not fit inside the window, or an unparsable sentence which does. For example (60) fits entirely inside the 6-word Local Association buffer, and hence the model incorrectly predicts it to be processable. Conversely, Marcus' 3-constituent window incorrectly predicts that (61) is a garden-path sentence.

(60) #The horse raced yesterday fell

(61) I know my aunt from Peoria died

Since these kinds of arguments have been made before, we turn to a new analysis of some non-garden path data. It has been argued that locality heuristics can account for the three kinds of preferences and grammaticality judgements in (63)–(64). First, Kimball (1973), Wanner (1980), and Gibson (1991) argue that iterative locality is what causes the adverb 'yesterday' to attach to the local verb 'died' rather than the distant verb 'said' or 'thought' in (62). Second, Kimball also argued that locality explained why (63a) could only mean that the job was attractive, not that the woman was attractive, and thus could not have the same sense as (63b). Finally, Kimball and Gibson have argued that iterative locality explains why examples like (64) are unacceptable, since the particle 'out' prefers to attach to the nearer predicate 'take' rather than 'figure'.

(62) a. Bill said John died yesterday.
b. Bill thought John died yesterday.

(63) a. The woman took the job that was attractive.
b. The woman that was attractive took the job.

(64) Joe figured that Susan wanted to take the train to New York out.

We argue that (63) cannot be explained by locality, and that no single locality heuristic can explain both (62) and (64). Consider the processing of restrictive relative clauses. Before Kimball's work, it was generally assumed that sentences such as (65a,b) (from Hankamer 1973) were simply ungrammatical:

(65) a. *A man$_i$ married my sister who$_i$ had castrated himself.

b. *I gave a kid$_i$ a banana who$_i$ was standing there looking hungry.

Kimball claimed that sentences like (65a,b) must be grammatical, because they were created by the same *Extraposition from NP* transformation that created (66b) from (66a). Since (Kimball claimed) (66b) was grammatical, (65a,b) must also be grammatical, and must only be ruled out for performance reasons. Thus the principle of *Right Association* would attach the phrase *who had castrated himself* to the noun *sister* instead of *man* in (65b), and thus (65) would be grammatical but unparsable.

(66) a. The woman that was attractive fell down

b. The woman fell down that was attractive. (*grammatical according to Kimball's theory*)

Our informants universally agreed, however, (although with no attempt at experimental confirmation) that (66b) is not at all grammatical, and so we star it for future reference.

(67) *The woman fell down that was attractive.

But (67) cannot be unacceptable because of either *windowing* locality or *iterative* locality. It cannot be accounted for by windowing locality since there are only two words (and only one constituent) between the relative clause and the nominal head, and no windowing theories predict such a small window. It cannot be unacceptable due to *iterative* locality, since there is no intervening nominal head; indeed there is no other nominal head in the sentence at all! [10]

The next phenomenon which is commonly cited as evidence for locality principles is the attachment of verbal particles to their head verbs. Kimball (1973) first presented the following examples, arguing that locality explains why (68a) is unacceptable, and why (68b) cannot have an interpretation in which the main verb of the sentence is *figure out*:

(68) a. Joe figured that Susan wanted to take the train to New York out.

b. Joe figured that Susan wanted to take the cat out.

For the windowing approaches like Local Association, the effect is caused by the correct head verb (*figured*) being too far from the particle (*out*). For the iterative approaches like Recency Preference, the effect is caused by the alternative head *take* which is in the way. Iterative or right-to-left locality fails as an explanation, however,

[10] One reviewer pointed out that making the relative clause heavier makes this sentence grammatical: "The woman fell down who had just stepped onto the moving platform". However, heavier phrases can often appear in sentence-final position in cases where non-heavy constituents are disallowed, and thus the acceptability of this sentence does not necessarily argue for the acceptability of (67).

because there are cases when a particle attachment is uninterpretable even if there is no possible intervening attachment point. For example, iterative models predict that a very long noun phrase without an embedded verb phrase should be interpretable, as there are no attachment points for verbal particles. However, (69a)–(69c) have no embedded verbs and yet are uninterpretable or at least quite difficult.

(69) a. *He threw the rotten apple from the tree behind our house out.
b. *I wrote that tedious problem set due Monday up.
c. *I called my friend, the one from New York, up.

It is possible that a kind of windowing locality might account for the phrasal verb examples in (70)–(69). However, the window cannot be based on number of words, as Fraser (1976) argued with the examples in (70). Note that (70a) includes a four-word noun phrase between the verb and particle and is uninterpretable. But (70b)–(70d), which include interrupting noun phrases with *five* words, are interpretable. Thus whatever the constraints may be on the placement of verb-particle objects, they are not statable in terms of constituent length.

(70) a. #I called *the man who left* up
b. He called *all of my best friends* up.
c. Won't you total *some of those larger figures* up.
d. Some charged *the adding machine fire-loss* off to experience.

It is possible that a kind of windowing locality might account for the phrasal verb examples in (69)–(70); note that in each of (69a)–(69c) a post-modified, phonologically heavy NP intervenes between the verb and the particle. In current work we are investigating whether a windowing locality approach together with the phonological weight criterion proposed by Zec and Inkelas (1990) (two phonological phrases) might explain the phrasal verb problem. However, significantly, even if this approach succeeds, this criterion cannot account for the different effects of the preferences in (62). The sentences in (69) are quite unacceptable; the preferences for rightmost attachment of the adverb *yesterday* in (62) are only slight preferences. Ford et al. (1982) give the following numbers for adverb attachment:

(71) Tom said Bill died yesterday.
a. Bill died yesterday. (70%)
b. Tom said it (that Bill died) yesterday. (30%)

Locality-type effects clearly play some role in a complete parsing model. But we have presented a paradox between iterative and windowing models of locality, arguing that it will be difficult to model the data with either of these approaches. In addition, we have noted that the serial parsing models which rely completely on heuristics like locality are unable to account elegantly for the delayed resolution of ambiguity needed in examples like (58).

5 Semantic disambiguation and a general model of interpretation

Until now we have ignored the role of semantics in disambiguation. However, a number of recent studies have shown that semantic context can reduce or eliminate the garden-path effect. Trueswell and Tanenhaus (1991), for example, show that garden path effects could be reduced by manipulating the tense of the clause. Crain and Steedman (1985) showed the effect of the semantic constraints a verb places on its arguments in examples like (72):

(72) a. #The teachers taught by the Berlitz method passed the test.
b. ?The children taught by the Berlitz method passed the test.

In current work with Srini Narayan we are investigating how our coherence model of disambiguation can be extended to deal with these semantic effects. Because we assume a sign-based or constraint-based theory of grammar, constructions are annotated with semantic information. One way of using semantic knowledge is to generalize the valence probabilities in the lexicon. Currently the probabilities refer to a solely syntactic specification of arguments. These probabilities could be extended to a function which assigns different probabilities to possible fillers of different semantic types, of the sort proposed by Resnik (1993). That is, the conceptual system of the language would be typed, and each valence slot of each predicate would be associated with a probability distribution over types. This would allow the valence probabilities for *teach*, for example, to distinguish *teachers* from *children* as prospective fillers. Burgess and Lund (1994) showed that computing psychological norms on similar simple thematic biases and using them to weight interpretations helped in modeling the semantic variation in garden-path sentences.

Another advantage of the probabilistic approach is that it could be extended from these kinds of 'grammatical' disambiguation examples to build a more complete theory of disambiguation. As Hirst (1986:111) noted, it is impossible to disambiguate sentences like (73a,b) without non-linguistic knowledge about "the relative aesthetics of factories and flora":

(73) a. The view from the window would be improved by the addition of a plant out there.
b. The view from the window would be destroyed by the addition of a plant out there.

A number of researchers have argued that a probabilistic model of abduction could be used to account for these sorts of non-linguistic disambiguation, making use of probabilistic real-world knowledge. (Charniak and Goldman 1988; Ng and Mooney 1990; Norvig and Wilensky 1990). Thus by including not only valence and construction probabilities, but also conceptual and non-linguistic probabilities, a broader probabilistic theory of interpretation could model a wide-range of data on access, sentence processing, and inference.

6 Problems and Future Work

The model as described suffers from a number of gaps and simplifying assumptions. The access algorithm assumes incorrectly that top-down and bottom-up evidence are

independent. We have not faced the difficult question of morphological processing nor addressed the recently burgeoning psycholinguistic literature on morphology, or modeled overload effects like center-embedding.

In addition, the model is currently unable to account for the kind of effects that have traditionally been modeled with spreading activation. These include intra-lexical semantic effects like the priming in 'The astronomer married the star' (Reder 1983). In addition, in an off-line experiment Gibbs et al. (1989) show effects of semantic priming in idiom processing, in which semantically decomposable idioms are processed faster than semantically non-decomposable idioms and than non-idiomatic control sentences. Our model could be fleshed out to deal with these kind of effects by adding a third kind of evidence to the access theory, some kind of semantic association evidence, in addition to top-down and bottom-up evidence.

On the other hand, these effects might also be handled on a different explanatory level of the theory, the level of activation. The model currently does not implement the time-course of construction activation. We have argued throughout the paper that access and disambiguation can be accounted for by ranking constructions by their posterior probability given the evidence; but we have not addressed implementing our model in the kind of activation or connectionist framework that is traditionally used to model low-level time-course (Feldman and Ballard 1982; Elman 1989; McClelland et al. 1989; MacDonald 1993). Producing an implementation in which the time-course of activation is proportional to posterior probability (rather than frequency) remains future work; it is possible that it is at this as-yet unaddressed level that semantic priming effects are best modeled. In addition, the work of Henderson (1994) in applying the connectionist paradigm of Shastri and Ajjanagadde (1993) to parsing suggests that effects like center-embedding which are commonly attributed to memory limitations could be modeled by a connectionist parser with certain memory-like limitations on variable binding. In addition, such an implementation level might allow us to incorporate the insights of competition-based approaches like that of Spivey-Knowlton and Tanenhaus (1995) or Stevenson (1993) in order to model the continuum of garden-path processing difficulty.

7 Conclusion

Traditional wisdom holds that a difficult problem can often be solved by *divide-and-conquer* methods; thus it has been argued that by dividing linguistic processing into modules for lexical, idiomatic, syntactic, and semantic processing, and orthogonally into separate models of access and disambiguation, we can eventually build a general theory of human language processing. Driven by experimental results, and resonating especially with the proposals of MacDonald (1993) and MacDonald et al. (1994), we have taken the opposite tack, proposing that a single probabilistic mechanism underlies the access and disambiguation of linguistic knowledge at every level, and demonstrating the model on psycholinguistic results at every level of linguistic structure, including lexical, idiomatic, and syntactic access and disambiguation, the interpretation of garden-path sentences, parsing preferences, and studies of gap-filling and other valence ambiguities.

Our theory makes another strong claim, regarding the use of probabilistic models

in linguistic theory. Generative linguistic theory has shied away from the use of probabilistic models since Chomsky's early arguments against Markov models of syntax. But the evidence we have presented here for the augmentation of each construction with probabilities, together with recent work which argues that probabilistic models are necessary to account for language change and learning, argues for a reanalysis of this position. Chomsky was correct in arguing against simple Markov models of syntax not because they were probabilistic, but because of their simplistic models of structure. We see probabilities not as replacements for structure, but as enrichments of structure; augmenting constructions with probabilities allows us to have the advantages of both structuralist and probabilistic models of language.

We hope this work also argues for holism at a different level, the level of academic disciplines. Building a cognitive model of parsing for a linguistic theory is necessarily an interdisciplinary enterprise. In particular, we have shown that models and metaphors from different disciplines of cognitive science can be used to solve problems in other sub-fields. For example, the psycholinguistic result that human processing of language is on-line was used to solve traditional computational complexity problems in parsing. Psycholinguistic results on the strong similarities in the processing of lexical, idiomatic, and syntactic structures were used to argue for sign-based models of linguistic structure like construction grammar, cognitive grammar, or HPSG. And finally, traditional computational algorithms like beam search are used to explain psychological results.

8 Acknowledgements

Many thanks to Jerry Feldman and Robert Wilensky for their significant contributions to and support of this work, and thanks to Charles Fillmore, Ted Gibson, James Greeno, Marti Hearst, Ron Kaplan, Jean-Pierre Koenig, George Lakoff, Ron Langacker, Clayton Lewis, Jim Martin, Don Mitchell, Srini Narayan, Terry Regier, Michael Spivey-Knowlton, Andreas Stolcke, Patrizia Tabossi, Paul Smolensky, Nigel Ward, and two anonymous reviewers for their contributions to this paper and earlier versions.

References

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

Altmann, G. T. M. and Steedman, M. J. (1988). Interaction with context during human sentence processing. *Cognition*, 30, 191–238.

Baker, J. K. (1979). Trainable grammars for speech recognition. In D. H. Klatt and J. J. Wolf (Eds.), *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, pages 547–550.

Barton, Jr., G. E., Berwick, R. C., and Ristad, E. S. (1987). *Computational Complexity and Natural Language*. MIT Press, Cambridge, MA.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the Development of Language*. Wiley, New York.

Bobrow, S. and Bell, S. (1973). On catching on to idiomatic expression. *Memory & Cognition*, 1, 343–346.

Boland, J. E. (1991). *The Use of Lexical Knowledge in Sentence Processing*. PhD thesis, University of Rochester, Rochester.

Boland, J. E., Tanenhaus, M. K., and Garnsey, S. M. (1990). Evidence for the immediate use of verb control information in sentence processing. *Journal of Memory and Language*, 29, 413–432.

J. Bresnan (Ed.) (1982). *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA.

Britt, A. (1991). *The role of referential uniqueness and argument structure in parsing prepositional phrases*. PhD thesis, University of Pittsburgh, Pittsburgh, PA. cited in Britt *et al*. (1993).

Britt, A., Perfetti, C. A., Garrod, S., and Rayner, K. (1992). Parsing in discourse: Context effects and their limits. *Journal of Memory and Language*, 31, 293–314.

Britt, M. A., Gabrys, G., and Perfetti, C. A. (1993). A restricted interactive model of parsing. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society (COGSCI-93)*, Boulder, CO.

Burgess, C. and Lund, K. (1994). Multiple constraints in syntactic ambiguity resolution: A connectionist account of psycholinguistic data. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society (COGSCI-94)*, Atlanta, GA.

Cacciari, C. and Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, 27, 668–683.

Cardie, C. and Lehnert, W. (1991). A cognitively plausible approach to understanding complex syntax. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, pages 117–124. Morgan Kaufmann.

Carlson, G. N. and Tanenhaus, M. K. (1987). Thematic roles and language comprehension. In W. Wilkins (Ed.), *Thematic relations*. Academic Press, San Diego.

Charniak, E. and Goldman, R. (1988). A logic for semantic interpretation. In *Proceedings of the 26th ACL*, Buffalo, NY.

Church, K. and Patil, R. (1982). Coping with syntactic ambiguity. *American Journal of Computational Linguistics*, 8, 139–149.

Clifton, Jr, C. (1993). The role of thematic roles in sentence processing. *Candian Journal of Psychology*, 47, 222–246.

Clifton, Jr, C. and Frazier, L. (1989). Comprehending sentences with long-distance dependencies. In G. N. Carlson and M. K. Tanenhaus (Eds.), *Linguistic structure in language processing*. Kluwer Academic, Dordrecht.

Clifton, Jr, C., Speer, S., and Abney, S. (1991). Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language*, 30, 251–271.

Connine, C., Ferreira, F., Jones, C., Clifton, C., and Frazier, L. (1984). Verb frame preference: Descriptive norms. *Journal of Psycholinguistic Research*, 13, 307–319.

Cottrell, G. W. (1985). Connectionist parsing. In *Proceedings of the 7th Annual Conference of the Cognitive Science Society*, pages 201–211, Irvine, CA.

Crain, S. and Fodor, J. D. (1985). How can grammars help parsers? In D. R. Dowty, L. Karttunen, and A. Zwicky (Eds.), *Natural language parsing*. Cambridge University Press, Cambridge.

Crain, S. and Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological syntax processor. In D. R. Dowty, L. Karttunen, and A. Zwicky (Eds.), *Natural language parsing*. Cambridge University Press, Cambridge.

Cutler, A. and Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–121.

Dalrymple, M. (1992). Categorial semantics for LFG. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 212–218, Nantes, France.

d'Arcais, G. B. F. (1993). The comprehension and semantic interpretation of idioms. In C. Cacciari and P. Tabossi (Eds.), *Idioms: Processing, Structure, and Interpretation*. Lawrence Erlbaum Associates, New Jersey.

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the Association for Computing Machinery*, 6, 451–455.

Elman, J. L. (1989). Structured representations and connectionist models. In *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, pages 17–25, University of Michigan, Ann Arbor, Mich.

Feldman, J. A. and Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, 6, 205–254.

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

Fillmore, C. J. (1986). Varieties of conditional sentences. In *ESCOL (Eastern States Conference on Linguistics)*, pages 163–182.

Fillmore, C. J. (1988). The mechanisms of "Construction Grammar". In *Proceedings of BLS 14*, pages 35–55, Berkeley, CA.

Fillmore, C. J., Kay, P., and O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64, 501–538.

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. *Linguistic Inquiry*, 9, 427–473.

Ford, M., Bresnan, J., and Kaplan, R. M. (1982). A competence-based theory of syntactic closure. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA.

Francis, W. N. and Kučera, H. (1982). *Frequency Analysis of English Usage*. Houghton Mifflin, Boston.

Fraser, B. (1976). *The Verb-Particle Combination in English*. Academic, New York.

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies*. PhD thesis, University of Connecticut, Connecticut. Dissertation distributed by the Indiana University Linguistics Club.

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and Performance XII*. Erlbaum, Hillsdale, NJ.

Frazier, L. and Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6, 291–295.

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

Fujisaki, T., Jelinek, F., Cocke, J., and Black, E. (1991). A probabilistic parsing method for sentence disambiguation. In M. Tomita (Ed.), *Current Issues in Parsing Technology*. Kluwer, Boston.

Garnsey, S. M., Tanenhaus, M. K., and Chapman, R. M. (1989). Evoked potentials and the study of sentence comprehension. *Journal of Psycholinguistic Research*, 18, 51–60.

Garrod, S. and Sanford, A. (1991). On the real-time character of interpretation during reading. *Language and Cognitive Processes*, 1, 43–59.

Gibbs, Jr, R. W. (1984). Literal meaning and psychological theory. *Cognitive Science*, 8, 275–304.

Gibbs, Jr, R. W., Nayak, N. P., and Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, 28, 576–593.

Gibson, E. (1991). *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

Gibson, E. and Pearlmutter, N. J. (1994). A corpus-based analysis of psycholinguistic constraints on preposition-phrase attachment. In *Perspectives on Sentence Processing*. Erlbaum, Hillsdale, NJ.

Glucksberg, S., Kreuz, R. J., and Rho, S. H. (1986). Context can constrain lexical access: Implications for modles of language comprehension. *Journal of Experimental Psychology: Learning Memory, and Cognition*, 12, 323–335.

Goldberg, A. E. (1991). A semantic account of resultatives. *Linguistic Analysis*, 21, 66–96.

Goldberg, A. E. (1992). The inherent semantics of argument stucture: the case of the English ditransitive construction. *Cognitive Linguistics*, 3, 37–74.

Gorrell, P. G. (1987). *Studies of Human Syntactic Processing: Ranked-Parallel Versus Serial Models*. PhD thesis, University of Connecticut, Connecticut.

Gorrell, P. G. (1989). Establishing the loci of serial and parallel effects in syntactic processing. *Journal of Psycholinguistic Research*, 18, 61–73.

Hankamer, J. (1973). Unacceptable ambiguity. *Linguistic Inquiry*, IV, 17–68.

Hearst, M. (1991). personal communication, noted in the Berkeley campus newspaper.

Henderson, J. (1994). *Description Based Parsing in a Connectionist Network*. PhD thesis, University of Pennsylvania, Philadelphia, PA.

Henderson, L. (1989). On mental representation of morphology and its diagnosis by measures of visual access speed. In W. Marslen-Wilson (Ed.), *Lexical Representation and Process*. MIT Press, Cambridge, MA.

Hill, A. A. (1970). Laymen, lexicographers, and linguists. *Language*, 46, 245–258.

Hirst, G. (1986). *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press, Cambridge.

Hirst, G. and Charniak, E. (1982). Word sense and case slot disambiguation. In *Proceedings of the Second National Conference on Artificial Intelligence*, pages 95–98.

Hobbs, J. R. and Bear, J. (1990). Two principles of parse preference. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, pages 162–167, Helsinki.

Hogaboam, T. W. and Perfetti, C. A. (1975). Lexical ambiguity and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, 14, 265–274.

Jelinek, F. and Lafferty, J. D. (1991). Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17, 315–323.

Joos, M. (1958). Review of Eric Hamp *A glossary of American technical linguistic usage* . *Language*, 34, 279–288.

Joos, M. (1972). Semantic axiom number one. *Language*, 48, 257–265.

Jurafsky, D. (1991). An on-line model of human sentence interpretation. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society (COGSCI-91)*, pages 449–454, Chicago.

Jurafsky, D. (1992b). An on-line computational model of human sentence interpretation. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, pages 302–308, San Jose, CA.

Jurafsky, D. (1992a). An on-line computational model of human sentence interpretation: A theory of the representation and use of linguistic knowledge. Technical Report 92/676, University of California at Berkeley dissertation, also available as Computer Science Division, Berkeley, CA.

Jurafsky, D., Fox, B., Morgan, N., and Stolcke, A. (1995a). NSF proposal: Probabilistic context-free grammars for spontaneous speech recognition. Submitted to CISE/IRI/Interactive Systems October 1995.

Jurafsky, D., Wooters, C., Tajchman, G., Segal, J., Stolcke, A., Fosler, E., and Morgan, N. (1995b). Using a stochastic context-free grammar as a language model for speech recognition. In *IEEE ICASSP-95*, pages 189–192.

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

Kay, M. (1973). The MIND system. In R. Rustin (Ed.), *Natural Language Processing*. Algorithmics Press, New York.

Kay, P. (1990). Even. *Linguistics and Philosophy*, 13, 59–216.

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

Koenig, J.-P. (1993). Linking constructions vs. linking rules: Evidence from French. In *Proceedings of BLS 19*, pages 217–231, Berkeley, CA.

Koenig, J.-P. (1994). *Lexical Underspecification and the Syntax/Semantics Interface*. PhD thesis, University of California, Berkeley, CA.

Koenig, J.-P. and Jurafsky, D. (1995). Type underspecification and on-line type construction in the lexicon. In *West Coast Conference on Formal Linguistics (WCCFL-94)*.

Krieger, H.-U. and Nerbonne, J. (1993). Feature-based inheritance networks for computational lexicons. In T. Briscoe, V. de Paiva, and A. Copestake (Eds.), *Inheritance, Defaults, and the Lexicon*. Cambridge University Press, Cambridge.

Kuno, S. (1965). The predictive analyzer and a path elimination technique. *Communications of the ACM*, 8, 453–462.

Kurtzman, H. S. (1985). *Studies in Syntactic Ambiguity Resolution*. PhD thesis, MIT, Cambridge, MA.

Kurtzman, H. S., Crawford, L. F., and Nychis-Florence, C. (1991). Locating WH- traces. In R. C. Berwick, S. P. Abney, and C. Tenny (Eds.), *Principle-Based Parsing: Computation and Psycholinguistics*. Kluwer, Dordrecht.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. University of Chicago Press, Chicago.

Lakoff, G. (1993). Cognitive phonology. In J. Goldsmith (Ed.), *The Last Phonological Rule*. University of Chicago Press, Chicago.

Lambrecht, K. (1995). The pragmatics of case. On the relationship between semantic, grammatical, and pragmatic roles in English and French. In M. Shibatani. and S. A. Thomspon (Eds.), *Essays in Semantics*. to appear.

Langacker, R. (1987). *Foundations of Cognitive Grammar I: Theoretical Prerequisites*. Stanford University Press, Stanford.

Luce, P. A., Pisoni, D. B., and Goldfinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing*. MIT Press, Cambridge, MA.

Luce, R. D. (1959). *Individual Choice Behavior*. Wiley, New York.

Lytinen, S. L. (1991). Semantics-first natural language processing. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, pages 111–116. Morgan Kaufmann.

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

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). Syntactic ambiguity resolution as lexical ambiguity resolution. In *Perspectives on Sentence Processing*. Erlbaum, Hillsdale, NJ.

Magerman, D. M. and Marcus, M. P. (1991). Pearl: A probabilistic chart parser. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany.

Manning, C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 235–242, Ohio State University, Columbus, Ohio.

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

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The Penn treebank. *Computational Linguistics*, 19, 313–330.

Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science*, 189, 226–228.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition*, 25, 71–102.

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing*. MIT Press, Cambridge, MA.

Marslen-Wilson, W., Brown, C. M., and Tyler, L. K. (1988). Lexical representations in spoken language comprehension. *Language and Cognitive Processes*, 3, 1–16.

Mathis, D. A. and Mozer, M. C. (1995). On the computational utility of consciousness. In G. Tesauro, D. S. Touretzky, and J. Alspector (Eds.), *Advances in Neural Information Processing Systems VII*, Cambridge, MA. MIT Press.

McCawley, J. D. (1989). The comparative conditional construction in English, German, and Chinese. In *Proceedings of BLS 14*, pages 176–187, Berkeley, CA.

McClelland, J. L. and Elman, J. L. (1986). Interactive processes in speech perception: The TRACE model. In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group (Eds.), *Parallel Distributed Processing Volume 2: Psychological and Biological Models*. MIT Press, Cambridge, MA.

McClelland, J. L., St. John, M., and Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, 4, 123–154.

McRoy, S. W. and Hirst, G. (1990). Race-based parsing and syntactic disambiguation. *Cognitive Science*, 14, 313–353.

Milne, R. W. (1982). Predicting garden path sentences. *Cognitive Science*, 6, 349–374.

Mitchell, D. C. (1989). Verb guidance and other lexical effects in parsing. *Language and Cognitive Processes*, 4, 123–154.

Ng, H. T. and Mooney, R. J. (1990). On the role of coherence in abductive explanation. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, pages 337–342. Morgan Kaufmann.

Norvig, P. (1987). *A Unified Theory of Inference for Text Understanding*. PhD thesis, University of California, Berkeley, CA. Available as University of California at Berkeley Computer Science Division Technical Report #87/339.

Norvig, P. (1988). Interpretation under ambiguity. In *Proceedings of BLS 14*, pages 188–201, Berkeley, CA.

Norvig, P. and Wilensky, R. (1990). A critical evaluation of commensurable abduction models for semantic interpretation. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, pages 225–230, Helsinki.

Orgun, O., Koenig, J.-P., and Jurafsky, D. (1995). Constraint-based morphology. submitted to *NELS-95*.

Osterhout, L. and Swinney, D. A. (1989). On the role of the simplicity heuristic in language processing: Evidence from structural and inferential processing. *Journal of Psycholinguistic Research*, 18, 553–562.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, San Mateo, Ca.

Perfetti, C. A. (1990). The cooperative language processors: Semantic influences in an autonomous syntax. In D. A. Balota, G. B. Flores d'Arcais, and K. Rayner (Eds.), *Comprehension Processes in Reading*. Lawrence Erlbaum, New Jersey.

Pollard, C. and Sag, I. A. (1987). *Information-Based Syntax and Semantics: Volume 1: Fundamentals*. University of Chicago Press, Chicago.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Potter, M. C. and Faulconer, B. A. (1979). Understanding noun phrases. *Journal of Verbal Learning and Verbal Behavior*, 18, 509–521.

Pritchett, B. (1988). Garden path phenomena and the grammatical basis of language processing. *Language*, 64, 539–576.

Pritchett, B. L. (1993). *Grammatical Competence and Parsing Performance*. University of Chicago Press, Chicago.

Reder, L. M. (1983). What kind of pitcher can a catcher fill? Effects of priming in sentence comprehension. *Journal of Verbal Learning and Verbal Behaviour*, 22, 189–202.

Resnik, P. (1992). Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 418–424, Nantes, France.

Resnik, P. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania. (Institute for Research in Cognitive Science report IRCS-93-42).

Riesbeck, C. K. and Schank, R. C. (1978). Comprehension by computer: Expectation-based analysis of sentences in context. In W. J. M. Levelt and G. B. F. d'Arcais (Eds.), *Studies in the perception of language*. Wiley, London.

Sag, I. A., Kaplan, R., Karttunen, L., Kay, M., Pollard, C., Shieber, S., and Zaenen, A. (1985). Unification and grammatical theory. In *Proceedings of the Fifth West Coast Conference on Formal Linguistics*.

Salasoo, A. and Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, 24, 210–231.

Schubert, L. K. (1984). On parsing preferences. In *Proceedings of the Tenth International Conference on Computational Linguistics*, pages 247–250.

Schubert, L. K. (1986). Are there preference trade-offs in attachment decisions? In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 601–605. Morgan Kaufmann.

Schvaneveldt, R. W., Meyer, D. E., and Becker, C. A. (1976). Lexical ambiguity, semantic context, and visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 243–256.

Selkirk, E. (1982). *The Syntax of Words*. MIT Press, Cambridge.

Shastri, L. and Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brian Sciences*, 16.

Shieber, S. M. (1983). Sentence disambiguation by a shift-reduce parsing technique. In *Proceedings of the 21st ACL*, pages 113–118, Cambridge, MA.

Shieber, S. M. (1985). Using restriction to extend parsing algorithms for complex-feature-based formalisms. In *Proceedings of the 23rd ACL*, pages 145–152, Chicago.

Simpson, G. B. and Burgess, C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance*, 11, 28–39.

Small, S. L. and Rieger, C. (1982). Parsing and comprehending with Word Experts. In W. G. Lehnert and M. H. Ringle (Eds.), *Strategies for Natural Language Processing*. Lawrence Erlbaum, New Jersey.

Spivey-Knowlton, M. (1994). Quantitative predictions from a constraint-based theory of syntactic ambiguity resolution. In M. Mozer, P. Smolensky, D. Touretzky, J. Elman, and A. Weigend (Eds.), *Proceedings of the 1993 Connectionist Models Summer School*, pages 130–137, Hillsdale, NJ. Erlbaum.

Spivey-Knowlton, M. and Sedivy, J. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition*, In press.

Spivey-Knowlton, M. and Tanenhaus, M. (1995). Discourse context and syntactic ambiguity resolution: Evidence from eye-movements during reading. submitted to the Journal of Experimental Psychology:LMC.

Spivey-Knowlton, M., Trueswell, J., and Tanenhaus, M. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology*, 47, 276–309.

Stevenson, S. (1993). A competition-based explanation of syntactic attachment preferences and garden path phenomena. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 266–273, Ohio State University, Columbus, Ohio.

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models*. PhD thesis, University of California, Berkeley, CA.

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21, 165–202.

Stolcke, A. and Omohundro, S. (1993). Hidden Markov Model induction by Bayesian model merging. In *Advances in Neural Information Processing Systems 5*. Morgan Kaufman, San Mateo, Ca.

Stowe, L. A. (1986). Evidence for on-line gap location. *Language and Cognitive Processes*, 1, 227–245.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18, 645–659.

Swinney, D. A. and Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, 18, 523–534.

Swinney, D. A. and Osterhout, L. (1990). Inference generation during auditory language comprehension. In A. C. Graesser and G. H. Bower (Eds.), *Inferences and Text Comprehension*. Academic Press.

Tabor, W. (1993). The gradual development of degree modifier *sort of*: A corpus proximity model. In K. Beals, G. Cooke, D. Kathman, K.-E. McCullough, S. Kita, and D. Testen (Eds.), *Proceedings of CLS 29*. University of Chicago. [Forthcoming.].

Tabossi, P., Spivey-Knowlton, M., McRae, K., and Tanenhaus, M. K. (1994). Semantic effects on syntactic ambiguity resolution: Evidence for a constraint-based resolution process. In *Attention and Performance XV*. Erlbaum, Hillsdale, NJ.

Tajchman, G., Jurafsky, D., and Fosler, E. (1995). Learning phonological rule probabilities from speech corpora with exploratory computational phonology. In *Proceedings of ACL-95*, pages 1–8, Cambridge, MA.

Tanenhaus, M. K., Boland, J. E., Garnsey, S. M., and Carlson, G. (1989). Lexical structures in parsing long-distance dependencies. *Journal of Psycholinguistic Research*, 18, 37–50.

Tanenhaus, M. K., Boland, J. E., Mauner, G., and Carlson, G. (1993). More on combinatory lexical information: Thematic structure in parsing and interpretation. In G. Altmann and R. Shillcock (Eds.), *Cognitive Models of Speech Processing*. Lawrence Erlbaum Associates, New Jersey.

Tanenhaus, M. K., Leiman, J. M., and Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior*, 18, 427–440.

Tanenhaus, M. K. and Lucas, M. M. (1987). Context effects in lexical processing. *Cognition*, 25, 213–234.

Tanenhaus, M. K., Stowe, L. A., and Carlson, G. (1985). The interaction of lexical expectation and pragmatics in parsing filler-gap constructions. In *Proceedings of the 7th Annual Conference of the Cognitive Science Society*, pages 361–365, Irvine, CA.

Taraban, R. and McClelland, J. L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language*, 27, 597–632.

Trueswell, J. C. and Tanenhaus, M. K. (1991). Tense, temporal context and syntactic ambiguity resolution. *Language and Cognitive Processes*, 6, 303–338.

Trueswell, J. C. and Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In *Perspectives on Sentence Processing*. Erlbaum, Hillsdale, NJ.

Tyler, L. K. (1984). The structure of the initial cohort: Evidence from gating. *Perception & Psychophysics*, 36, 417–427.

Tyler, L. K. and Marslen-Wilson, W. (1982). Speech comprehension processes. In J. Mehler, E. C. T. Walker, and M. Garrett (Eds.), *Perspectives on mental representation.* Erlbaum, Hillsdale, NJ.

Ulvestad, B. (1960). On the use of transitional probability estimates in programming for mechanical translation. *Statistical Methods in Linguistics*, 1, 24–40.

van der Linden, E.-J. (1992). Incremental processing and the hierarchical lexicon. *Computational Linguistics*, 18, 219–238.

van der Linden, E.-J. and Kraaij, W. (1990). Ambiguity resolution and the retrieval of idioms: two approaches. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, pages 245–250, Helsinki.

van Noord, G. (1991). Head corner parsing for discontinuous constituency. In *Proceedings of the 29th ACL*, pages 114–121.

Wanner, E. (1980). The ATN and the Sausage Machine: Which one is baloney. *Cognition*, 8, 209–226.

Whittemore, G., Ferrara, K., and Brunner, H. (1990). Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. In *Proceedings of the 28th ACL*, pages 23–30, Pittsburgh, PA.

Wilensky, R. and Arens, Y. (1980). PHRAN: A knowledge-based natural language understander. In *Proceedings of the 18th ACL*, Philadelphia.

Wilks, Y. (1975a). An intelligent analyzer and understander of English. *Communications of the ACM*, 18, 264–274.

Wilks, Y. (1975b). Preference semantics. In E. L. Keenan (Ed.), *The Formal Semantics of Natural Language*. Cambridge Univ. Press, Cambridge.

Wilks, Y., Huang, X., and Fass, D. (1985). Syntax, preference and right attachment. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 779–784.

Wright, B. and Garrett, M. (1984). Lexical decision in sentences: Effects of syntactic structure. *Memory & Cognition*, 12, 31–45.

Zec, D. and Inkelas, S. (1990). Prosodically constrained syntax. In S. Inkelas and D. Zec (Eds.), *The Phonology-Syntax Connection*. University of Chicago, Chicago.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, 32, 25–64.