Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cognition 88 (2003) 133–169
COGNITION
www.elsevier.com/locate/cognit

Learning first-pass structural attachment preferences with dynamic grammars and recursive neural networks

Patrick Sturt[a,*], Fabrizio Costa[b], Vincenzo Lombardo[c], Paolo Frasconi[b]

[a]*Department of Psychology, University of Glasgow, 58 Hillhead Street, Glasgow G12 8QB, UK*
[b]*Department of Systems and Computer Science, University of Florence, Via Santa Marta 3, 50139 Firenze, Italy*
[c]*Dipartimento di Informatica, Università di Torino, Corso Svizzera 185, 10149 Torino, Italy*

Received 8 March 2001; received in revised form 21 May 2002; accepted 3 December 2002

Abstract

One of the central problems in the study of human language processing is ambiguity resolution: how do people resolve the extremely pervasive ambiguity of the language they encounter? One possible answer to this question is suggested by experience-based models, which claim that people typically resolve ambiguities in a way which has been successful in the past. In order to determine the course of action that has been "successful in the past" when faced with some ambiguity, it is necessary to generalize over past experience. In this paper, we will present a computational experience-based model, which learns to generalize over linguistic experience from exposure to syntactic structures in a corpus. The model is a hybrid system, which uses symbolic grammars to build and represent syntactic structures, and neural networks to rank these structures on the basis of its experience. We use a dynamic grammar, which provides a very tight correspondence between grammatical derivations and incremental processing, and recursive neural networks, which are able to deal with the complex hierarchical structures produced by the grammar. We demonstrate that the model reproduces a number of the structural preferences found in the experimental psycholinguistics literature, and also performs well on unrestricted text.

Keywords: Connectionism; Parsing; Hybrid models; Ambiguity resolution

Natural language is vastly ambiguous, and our apparently effortless ability to understand it is one of the central problems of modern cognitive science. A number of current

* Corresponding author. Tel.: +44-141-330-3606; fax: +44-141-330-4606.
E-mail address: patrick@psy.gla.ac.uk (P. Sturt).

0010-0277/02/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0010-0277(03)00026-X

theories of human parsing claim that a central role is played by linguistic experience in ambiguity resolution. According to such theories, ambiguities are preferentially resolved in a way that has been successful most frequently in the past. Experience-based mechanisms could work at a number of different levels of processing, from the low levels associated with lexical access to higher-level discourse processing. In this paper, however, we will be concerned with purely syntactic factors in experience-based ambiguity resolution. As we will see below, there is converging agreement that frequencies of syntactic structures play an important role in determining ambiguity resolution preferences in human parsing. Below, we propose a hybrid model of a purely syntactic experience-based ambiguity resolution mechanism which involves a novel machine learning architecture, recursive neural networks (Frasconi, Gori, & Sperduti, 1998). The goal of the approach is to create a model which is successful both in reproducing well-known experimental results in psycholinguistics and in dealing with input drawn from unrestricted text. In this paper, we show how the model can be trained given a small sample from a parsed corpus (treebank), leading to good disambiguation performance both on unrestricted text, and on a collection of examples from the psycholinguistic experimental literature. The performance is particularly encouraging given the small size of the training set and the extremely pervasive ambiguity that has to be considered when working on a realistically large scale.

1. Purely syntactic frequency effects

There have been at least two proposals in which purely syntax-based frequencies have been claimed to play a role in syntactic disambiguation. According to the *tuning hypothesis* (Mitchell, Cuetos, Corley, & Brysbaert, 1995) "early parsing choices can be determined by high-level statistical regularities of the language" (p. 485). That is, the initial disambiguation choice is made on the basis of the frequency of certain types of structures which have been found in past experience. Consider the following sentence, which is similar to the sentence type employed by Cuetos and Mitchell (1988):[1]

1. The servant of the actress who was on the balcony died.

Cuetos and Mitchell (1988) found evidence that English speakers preferred to attach the ambiguous relative clause to the more recent of the two possible host nouns ("low attachment"), supporting the interpretation that the actress was on the balcony. For the equivalent Spanish sentence on the other hand, native speakers preferred to attach the relative clause to the less recent noun ("high attachment"), supporting an interpretation in which the *servant* was on the balcony. In later work, Mitchell et al. (1995) proposed that this difference reflected differences in the structural frequencies of the two languages. In English, configurations of the form $[_{NP} \ldots [_{PP} P [_{NP} \ldots RC]]]$ are more frequent than $[_{NP} \ldots [_{PP} P NP] RC]$, and initial investigations suggested that the reverse is true for Spanish. Moreover, it has also been demonstrated that preferences for sentences such as (1) can be modified by exposure to appropriately disambiguated examples over a two-week period (Cuetos, Mitchell, & Corley, 1996), again supporting an experience-based model. It

[1] Here we put the complex noun phrase in subject position as in Traxler, Pickering, and Clifton (1998).

should be noted that the Tuning Hypothesis is concerned with purely syntactic frequencies in the initial ambiguity choice. Clearly, ambiguity resolution also involves non-syntactic factors, such as lexical frequency and semantic plausibility. However, according to the Tuning Hypothesis, these non-syntactic factors play a role in later processes, which filter the output of the first-pass ambiguity resolution mechanism. In this respect, the Tuning Hypothesis resembles the Garden Path model (Frazier, 1987).

Mitchell et al. (1995) offer empirical support for the claim that there is a purely configurational, non-lexical component to disambiguation preferences. In one study, they had English-speaking participants complete sentence fragments like *The satirist ridiculed the lawyer of the firm wh…* At a subsequent session, three weeks later, the same participants were given parallel materials, with the head-noun positions of the critical noun phrases reversed (e.g. *The satirist ridiculed the firm of the lawyer wh…*). They found that, in addition to a reliable low attachment preference across sessions, individual participants' preferences for low or high attachment correlated positively across the two sessions. Thus, in this study, the consistent preference was to attach the relative clause to a particular *configurational position*, rather than to use the relative clause to modify a particular lexical item – if the preference had been purely lexical, participants' preferences should have reversed from one session to the other, because of the reversal of the position of the head nouns.

The second theory in which purely syntactic frequencies are hypothesized to play a major role is the multiple constraints framework (McRae, Spivey-Knowlton, & Tanenhaus, 1998; Spivey & Tanenhaus, 1998). In the model proposed by McRae et al. (1998), syntactic ambiguity resolution is achieved using a competition mechanism, in which a number of constraints simultaneously compete with each another to influence the relative activations of the alternative readings of the ambiguity. Consider the following sentence, for example:

2. The cook arrested by the detective was guilty of taking bribes.

The attachment of the word *arrested* is temporarily ambiguous. It could be the main verb of the sentence (with *the cook* as its subject), or it could be part of a reduced relative clause, which is the attachment that turns out to be correct in (2). McRae et al. (1998) claim that a number of constraints are active from the earliest stages of processing in determining the preferred attachment of the verb. Some of these constraints are lexical, such as the plausibility of *the cook* as the agent of *arrested* relative to its plausibility as the patient of *arrested*. Another lexical constraint is the frequency with which the verb appears in past tense form compared with how often it appears as a past participle. Each of these constraints favours the reduced relative and main clause readings to a different degree, depending on the actual lexical content of the verb and the head of the preceding noun phrase. However, another proposed constraint, which McRae et al. call *Main Clause Bias* is purely syntactic, and relates to the frequency of the two alternative syntactic structures without regard to specific lexical content. As they point out, 'A sentence-initial sequence "noun phrase verbed" is typically the beginning of a main clause' (see also Bever, 1970). In other words, the constraint favours the main clause reading in any sentence which matches the sequence, regardless of the precise lexical form of the verb or the preceding

head noun. Clearly, this Main Clause Bias constraint plays a very similar role to the Tuning mechanism, which resolves ambiguities according to high-level (syntactic) statistical regularities in the language. The main difference between the approaches is in the architectural relation between the syntactic and lexical constraints. In the Tuning framework, structural preferences based on purely syntactic frequencies have a privileged place in the ambiguity resolution process, as it is these preferences which determine the initial resolution of a syntactic ambiguity, while lexical and other non-syntactic constraints only operate in a later filtering stage. In contrast, in the multiple constraints framework, purely structural frequencies act simultaneously with lexical constraints, and do not play any privileged architectural role.

The fact that purely syntactic frequencies play a role in both accounts discussed above brings us to the question of how to count these syntactic frequencies, and how to model their application in ambiguity resolution during incremental processing. In the remainder of this paper, we will describe the problems involved in building such a model and our solution to them, which will also consider the case of processing unrestricted text.

2. The problem of tree generalization

As we have seen above, purely syntactic frequencies have been used to model ambiguity resolution in both the Tuning and Multiple Constraints frameworks. Researchers have proposed to estimate the frequencies of the structural alternatives of particular ambiguities by counting examples in corpora. This raises two problems. The first problem is that of exactly what one should be counting. Take the relative clause example discussed in the Tuning approach. The two possible attachments can be (partially) represented as follows:

3. (a) **low attachment**
   $[_{NP1}$ the servant $[_{PP}$ of $[_{NP2}$ the actress $[_{SBAR}$ who was on the balcony]]]]
   (b) **high attachment**
   $[_{NP1}$ the servant $[_{PP}$ of $[_{NP2}$ the actress]] $[_{SBAR}$ who was on the balcony]]

In order to estimate frequency-based structural preferences directly from corpora, it is necessary to search corpora for all the sentences which match the two configurations given above. The problem is how to count the structures, that is, how to determine whether some given configuration in the corpus "matches" the target configuration (this is also known as the "grain" problem; Mitchell et al., 1995). There are many different possible ways of counting even if we confine ourselves to purely non-lexical definitions. In the Tuning Hypothesis, the defining structure is usually assumed to be NP-PP-RC, where NP dominates PP, and RC (the relative clause) could be immediately dominated either by the first NP, or by the second NP (which is embedded inside the PP). This scheme still leaves open a number of possibilities, discussed by Mitchell et al. (1995). For example, should the internal structure of each noun phrase match the target exactly (e.g. should it have one determiner, one noun etc.)? If not, can the noun phrases be arbitrarily complex? Similar remarks can be given for the reduced relative/main clause example. McRae et al. counted

the frequencies of the two alternatives for verbs following noun phrases in sentence initial position, but there are many other ways in which the structures could have been counted. For example, the internal structure of the noun phrase could have been considered, or the structures could have been counted in positions other than sentence initial positions. Clearly, it is desirable to have a principled and generally applicable method of counting structures in a corpus.

The second problem that has to be faced is that of sparse data. The more detailed the matching scheme for finding target configurations, the less likely one is to find examples in any given corpus. Sometimes, there may be no exact matches for some configuration, and in such cases, it is necessary to generalize on the basis of past experience of non-identical but similar structures. The method of generalization should be adequate and correct, in the sense that it should correspond to the actual structural preference that would be employed by a human when faced with the same input.

In this paper, we propose a novel approach for addressing these problems. The model has a hybrid architecture (see also Stevenson, 1994; Vosse & Kempen, 2000). The symbolic component of the model involves a dynamic grammar, which provides a very close link between competence and performance for incremental processing. The numerical component of the model involves a recursive neural network architecture, which is used to rank the alternative structures generated by the symbolic component, according to experience-based structural preferences. The hybrid model differs both from purely symbolic models, which are not suitable for investigations of experience-based preferences, and from purely connectionist models, which typically have difficulty in scaling up beyond small-scale linguistic domains (though see Rohde (2002) for a recent proposal of a purely connectionist model that captures a wide range of human sentence-processing data on an impressively large subset of English).

There is no explicit matching scheme for counting configurations in the corpus, because the network does not have the representational capacity to store all the trees to which it is exposed. The network learns to recognize the *abstract* features of trees that can be used to make comparisons with similar but possibly non-identical examples previously encountered during training. Thus, the model does not work by attempting to match configurations against a store of previously seen examples. The question of which dimensions of variation the network considers in making its decision is an empirical one. The results of this paper will clearly show that the network pays attention to at least two psychologically interesting dimensions; it is sensitive to the height of attachment of various types of phrases to partial trees, and it is also sensitive to the syntactic complexity of attachments.

Clearly, in order to be complete, the processing model would also need to treat lexical, semantic and other factors which are known to affect human ambiguity resolution. The model which we present in this paper uses pure phrase-structure trees, without lexical content, and we therefore cannot consider these other relevant factors. However, we note that, in principle, the configurational-based model that we present here could be combined with non-syntactic mechanisms in a number of ways. For example, it could form the initial ambiguity resolution component of a Tuning-based model, whose output is later filtered with reference to lexical, semantic and other factors. The model could also be usefully employed within the multiple constraints framework. Our technique can be used to generate, on a word-by-word basis, the syntactic alternatives that are evaluated by the

constraints. Recent constraint-based implementations (McRae et al., 1998; Spivey & Tanenhaus, 1998) have tended to lack a mechanism for generating the syntactic alternatives that are evaluated by the constraints. Moreover, the learning method that we present here could be used to estimate the preferences expressed by the purely structure-based constraint (as in the *Main Clause Bias* constraint described above).

In any case, the advantages of the present model are firstly that it uses a general, non-stipulatory method for matching structures to past experience and secondly that it will provide frequency estimates for any given ambiguity, including ones that it has not seen before, thus extending its capabilities beyond well studied ambiguities. In the future, it may also be possible to extend the model described here to allow lexical items to be included in the hybrid representations of trees, thus dealing with lexical and configurational phenomena in a unitary architecture.

The model we present here predicts first-pass attachment preferences. In the overall architecture, the model described here is assumed to act as an oracle for attachment decisions in an incremental parser. We assume that this parser analyzes the sentence from left to right and maintains a fully connected partial structure as each word has been read. In the next section we sketch the incremental processing model that we are assuming. Then we describe the recursive neural network architecture and how we tackle the problem of ambiguity resolution. Finally, we illustrate the behaviour of the model in processing real text, as well as in a number of well-known examples drawn from the psycholinguistic literature.

3. Incrementality, connectedness and dynamic grammars

There is widespread agreement in the psycholinguistic community that human sentence processing is highly incremental. Thus, Steedman (1989, Chap. 16) claimed that processing proceeds left-to-right, and that parsing commitments are made as each word is processed. This view is supported by a large body of experimental evidence, showing that semantic interpretations of ambiguous attachments are available before disambiguating information appears (Pickering & Traxler, 1998), that syntactic attachment decisions are made before the head of the relevant phrase appears (Bader & Lasser, 1994; Kamide & Mitchell, 1999), and that people are able to "shadow" speech at extremely short latencies (Marslen-Wilson, 1973). In algorithmic terms, the simplest way to model incrementality is to insist that each word in the input is immediately attached to a fully connected syntactic representation (Lombardo & Sturt, 2002; Stabler, 1994, Chap. 13). This implies that the syntactic representation expands as each word is read and incorporated into it. In this paper, we will rely on such a conceptualization of incremental processing. Before we proceed, however, we will clarify some terminology.

3.1. The idealized incremental parsing model

We assume that the target syntactic representation of a sentence is a phrase structure tree, which remains fully connected as it is built up during word-by-word incremental processing. Given the fully-connected conceptualization of incrementality, the phrase structure tree will grow as each word is input. The partial tree that is built up during

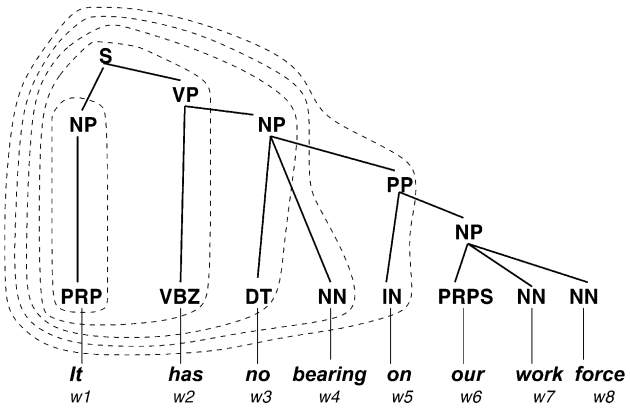

Fig. 1. The tree for a sentence in the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993). Incremental trees for the first five words are enclosed in dotted loops.

processing will be called the *Incremental Tree*. As an example, consider Fig. 1. As each word is read, the incremental tree is expanded, as shown for the first five words ($w_1 \ldots w_5$) in the figure. The incremental tree that is created when w_i is attached will be referred to as T_i. When a new word is read, it is connected to the previous incremental tree via what we will call a *Connection Path*. Intuitively, the connection path for some word w_i consists of the portion of the new incremental tree T_i that is not part of the previous incremental tree T_{i-1}. More precisely, the connection path for word w_i consists of all the branches of the tree that are in T_i but not in T_{i-1}, plus all the nodes of the tree that are directly connected to these branches. The node that is in common between T_i and T_{i-1} will be referred to as the *anchor* for the connection path. Intuitively, the anchor is the attachment site for the new word in the left context. The *foot* node of a connection path is the preterminal part-of-speech category that immediately dominates the new word. For present purposes (i.e. for determining the influence of purely syntactic structures) we assume that the connection paths do not include the actual lexical items. For example, the connection paths for the first five words of the sentence above are shown in Fig. 2. To summarize, an idealized incremental parsing of a sentence $S = w_1, w_2, \ldots w_n$ proceeds through a sequence of incremental trees, $T_1, T_2, \ldots T_n$. Given a word w_i and an incremental tree T_{i-1} (its left context), the task is to find a connection path from w_i to T_{i-1} to create T_i.

Attachment ambiguity exists when for some word w_i, and its left context T_{i-1}, there exist more than one possible T_i. This can happen because there is more than one possible connection path, or because there is more than one anchor in T_{i-1} to which some connection path could attach, or there could be multiple connection paths and multiple anchors. The ambiguity resolution problem is schematically illustrated in Fig. 3. This figure shows a word, w_i, which could attach to any one of three possible anchors in the incremental tree T_{i-1}, and each of these attachment sites could be reached via two possible connection paths. Thus, there are six different ways of attaching w_i to T_{i-1} to create T_i. The ambiguity resolution problem is to find which of these is the correct one.

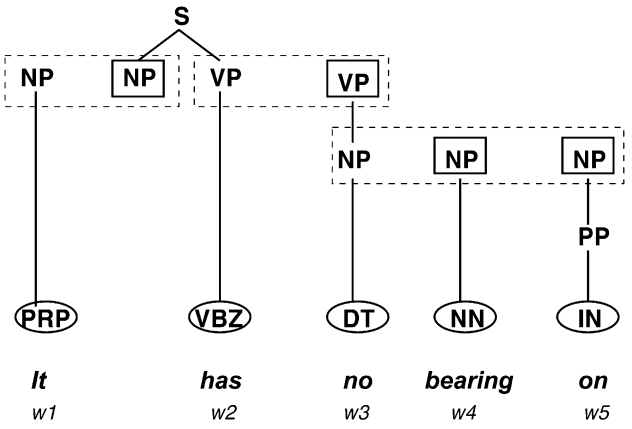

Fig. 2. Connection paths for the first five words of the sentence in Fig. 1. Anchors are enclosed in boxes, and foot nodes are enclosed in ovals. Dotted boxes enclose the nodes which will be equated in the parsing process.

3.2. Dynamic grammar

A dynamic grammar defines linguistic well-formedness in terms of states, and transitions between states (Milward, 1994). In terms of the present model, the incremental trees can be seen as states in a left-to-right incremental parse. The connection paths define the possible transitions between states, and can thus be seen to constitute a dynamic grammar. For example, a dynamic grammar for a very tiny fragment of English consists of the five

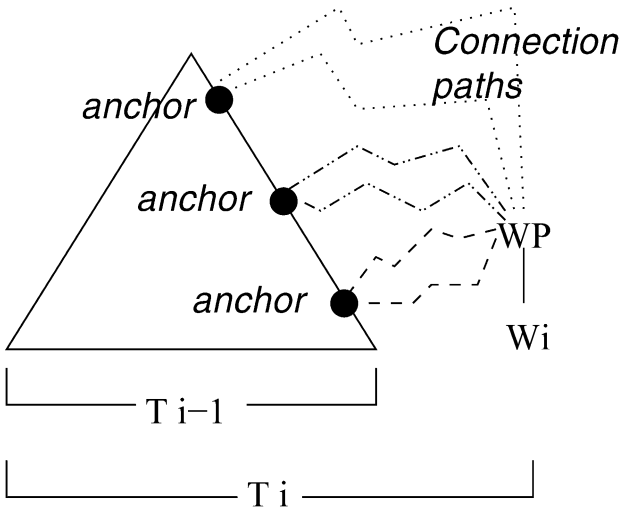

Fig. 3. The ambiguity resolution problem. There may be many possible connection paths that could connect a new word w_i to T_{i-1}, and each of these may be attachable to multiple possible anchors.

connection paths depicted in Fig. 2. In the model which we present in this paper, we assume two constraints on connection paths:

For any word w_i and an incremental tree T_{i-1}

(1) the anchor of the connection path matches the category of a node on the right frontier of T_{i-1} (where the *right frontier* of an incremental tree is the set of nodes in that tree which precede no other node);
(2) the foot of the connection path matches the part-of-speech tag of w_i.

In Lombardo and Sturt (2002), an algorithm is presented which can extract a set of connection paths from a treebank (parsed corpus). The algorithm works by taking each tree in the treebank corpus, and "simulating" an incremental parse of the corresponding sentence. The "simulation" is done by taking each word of the sentence in left-to-right order, and marking those branches of the tree that would allow that word to be attached to the previous incremental tree, in a fully connected manner. The portion of the tree that is marked at each step corresponds to the connection path for the current word. For example, in the tree shown in Fig. 1, the branches that would be marked for each of the first five words are indicated by the dotted loops: for each word w_i shown in the figure, the relevant branches are those that lie outside the dotted loop for the previous word w_{i-1}, but inside the dotted loop for the current word w_i. Simulations reported by Lombardo and Sturt (2002) indicate that the connection paths required to create a dynamic grammar for a treebank are typically very simple and predictable. For example, there are strict limitations on the number of distinct "headless projections" (nodes whose heads have not yet been read in the input) that need to be included in connection paths. This gives some support to the notion that it would be feasible to parse using these connection paths directly as the grammar.

Dynamic grammars provide a simultaneous definition of syntactic well-formedness and possible parsing actions, and therefore have the advantage of simplifying the relation between competence and performance. Recent work has shown that defining grammars in terms of left-to-right structure building can solve a number of problems in theoretical syntax (Kempson, Meyer-Viol, & Gabbay, 2000; Phillips, 1996; Tugwell, 1998), and dynamic grammars related to ours have also been used to model language acquisition (Fodor, 1998). For the task that we are concerned with in this paper, i.e. learning structural preferences in parsing, the advantage of dynamic grammars is that the incremental trees can be seen as a realistic approximation of the left contexts that are hypothesized by psycholinguists to exist when parsing decisions are made.

4. Experience-based models of structural preferences

The disambiguation problem for the attachment of some word w_i is to find the correct connection path and anchor, given some incremental tree T_{i-1}. This problem is equivalent to that of finding the correct T_i produced by attaching the correct connection path to the correct anchor in T_{i-1}. The learning problem for an experience-based model is therefore to select the best T_i out of all the possibilities, given the past experience. In other words, we

need a method for judging trees against linguistic experience. In this paper, we will describe a learning method for this task based on recursive neural networks. Before we describe this, however, we will set the context by describing a popular alternative approach, based on probabilistic context-free grammars.

4.1. Probabilistic context-free grammars

The utility of probabilistic context-free grammars for psycholinguistic modelling was first demonstrated by Jurafsky (1996), where the formalism was used for predicting human preferences in syntactic and lexical category disambiguation. The general idea of this approach is that the probability of a tree (or a partial tree being built up during incremental processing) can be estimated by combining (usually multiplying) the probabilities of each context-free rule that is used in its derivation. The actual rule probabilities can be estimated by taking the frequency of each rule's use in a treebank corpus, and normalizing over the left-hand sides of the rules. For example, imagine that we want to estimate the probability of the rule NP → D N, which allows a noun phrase to be expanded into a determiner followed by a noun. This rule describes a small portion of a tree one branch deep, consisting of one NP node with two daughters (D and N). If this small fragment of structure appears 800 times in a treebank corpus, and all the other one-branch-deep fragments rooted in NP appear for a total of 200 times, then we can normalize to give a probability of 0.8 to the rule NP → D N. In the model presented by Jurafsky (1996), context-free rule probabilities like these are combined with valence probabilities, based on frequencies of subcategorization frames. The result is that disambiguation preferences can be modelled by estimating and comparing the probabilities of the two (or more) lexicalized trees that represent the different readings of the ambiguity, and choosing the tree with the highest probability as the preferred alternative. For example, Fig. 4, taken from Jurafsky's paper, illustrates how the model is used to disambiguate a lexical category ambiguity, where the various rule probabilities are multiplied to gain probabilities for partial trees (though for full details, see Jurafsky, 1996). Multiplying the respective rule probabilities shows that the probability for tree (a) on the left of the figure is higher than the probability for tree (b) on the right of the figure. Therefore, Jurafsky's model predicts that, in a sentence like *The complex houses married and single students*, the initially preferred reading is one in which *complex* is an adjective and *houses* is a noun. In this case, the main reason for the difference in overall tree probabilities is the difference in

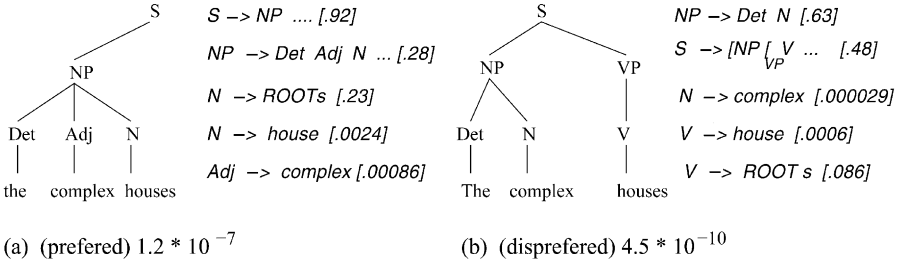

Fig. 4. Estimating tree probabilities (from Jurafsky, 1996, p. 170).

probabilities between the rules which define the lexical categories of the words *house* and *complex*. Because this analysis is incompatible with the sentence as a whole, the model therefore predicts a garden path effect for this sentence.

More recent work has shown that the use of probabilistic context-free parsing techniques to model incremental processing can be scaled-up effectively to deal with naturally occurring text (Brants & Crocker, 2000; Roark & Johnson, 1999). Such large-scale systems, in combination with Markovian learning methods, can also be used to model probabilistic preferences in human parsing (Crocker & Brants, 2000).

If we return to consider the problem of sparse data in estimating the probabilities of trees, we can see that probabilistic context-free grammars offer one possible solution; even though the frequency of some tree may not be known (i.e. the tree as a whole has not appeared previously in the model's "experience"), the tree can be decomposed into smaller, local chunks (i.e. context-free rules), each of which will often correspond to a context-free rule with known frequency.

This means that context-free grammars are most suitable for modelling preferences which can be characterized in terms of local tree configurations. Thus, in the example discussed above (Fig. 4), the crucial choice is between different sets of context-free rules that assign parts of speech to the words *complex* and *houses*. In many cases, it is also possible to characterize syntactic attachment ambiguities in terms of a choice between local context-free rules. For example, Jurafsky (1996) discusses how the choice of context-free rules interacts with lexical preferences to predict prepositional phrase attachment preferences in sentences like *They discuss the dogs on the beach*, where the prepositional phrase can attach either to the noun phrase or the verb phrase. The crucial difference between the two interpretations can be captured by the fact that the verb phrase attachment and the noun phrase attachment require different sets of context-free rules, and these rules have different probabilities. These different rule probabilities then interact with different valence probabilities for verbs like *discuss* or *keep*, to predict different attachment preferences based on both lexical and syntactic information.

However, many of the preferences discussed in the sentence-processing literature are difficult to characterize in terms of a choice between different context-free rules, and are better thought of in terms of larger global configurations. Recall that experience-based models have been discussed extensively in relation to examples like 1, repeated below:

1. The servant of the actress who was on the balcony died.

In many context-free grammars, the trees representing the two readings of this ambiguity would employ exactly the same set of context-free rules in the derivation. This is because the crucial ambiguity involves a choice of *where* in the tree (or when in the derivation) to apply the rule that is used to attach a relative clause to a noun phrase, rather than a choice between two different rules. If both trees employ the same rules, then the product of the rule probabilities should not differ between the trees representing the two candidate attachments, and thus in many systems based on probabilistic context-free grammars, it would be difficult to capture the relevant differences in frequency between the two constructions.

Intuitively, what we need is a way of representing the relative positions of different

attachment sites in a tree. In other words, we need a way of representing *global* information about an incremental tree, and not just the *local* information that can be defined in terms of combinations of context-free rules.

One way of representing global preferences in an experience-based system would be to count statistics over larger structured objects than those that are available in context-free grammars. As we have seen above, this is the approach that was suggested by Mitchell et al. (1995). Mitchell et al. suggest that humans keep frequency records of the various defining configurations for structural ambiguities (such as the configuration that defines the two-site relative clause ambiguity discussed above, for example). However, as we have discussed above, dealing with larger structures leads to problems with sparse data, since the larger the target structure is, the less often it will appear in any given corpus.

The approach that we describe in this paper can be seen as a method for collecting statistics over large configurations, but one which does not rely on counting occurrences of specific target configurations. The model forms its own (distributed) representations of the abstract structural features that are relevant to ambiguity resolution. We will see later that this gives the model the ability to generalize so that it can propose systematic preferences to ambiguities of a type that have been unseen in training.

4.2. Recursive neural networks

Our approach will rely on *Recursive Neural Networks* (Frasconi et al., 1998; Goller & Kuechler, 1996; Sperduti & Starita, 1997) (henceforth RNNs), which are capable of learning to classify hierarchical data structures, such as the incremental trees which we employ in this paper. We will see below that, unlike the probabilistic context-free grammar formalism, this allows us to represent global information about trees, such as the relative position of different attachment sites. It also provides a natural way to generalize to unseen examples.

The task of the model is to take any given word w_i and incremental tree T_{i-1}, and to rank the candidate incremental trees that can be produced by attaching w_i to T_{i-1} using the dynamic grammar. The highest ranked tree will be chosen as the preferred alternative.

Before discussing the learning method, we will outline the general architecture of the RNN model. An RNN consists of two major components: the recursive network proper, and an output network. The recursive network proper is a feedforward network which is "unfolded" according to the topology of the input tree, as illustrated in Fig. 5. The figure shows the recursive and output networks, and an example "unfolding" of the network for a simple input tree. The *unfolding* consists of a replication of the recursive network at each node in the tree. The *output* network processes the output of the root node of the tree.

The recursive network associated with each node v in the tree is a simple feedforward network, whose output is a *hidden state vector* $X(v) \in R^n$ (a vector of length n of real numbers), which, following training, will encode the features of the subtree dominated by v which are relevant to the task (in our case, a disambiguation task). The dimension n must be large enough to give sufficient expressive power to the network. In the experiments reported in this paper, we used a vector of length 20, allowing the network to encode trees in a 20-dimensional space. The state vector associated with a node v is computed by a state transition function, learned through training, which combines the

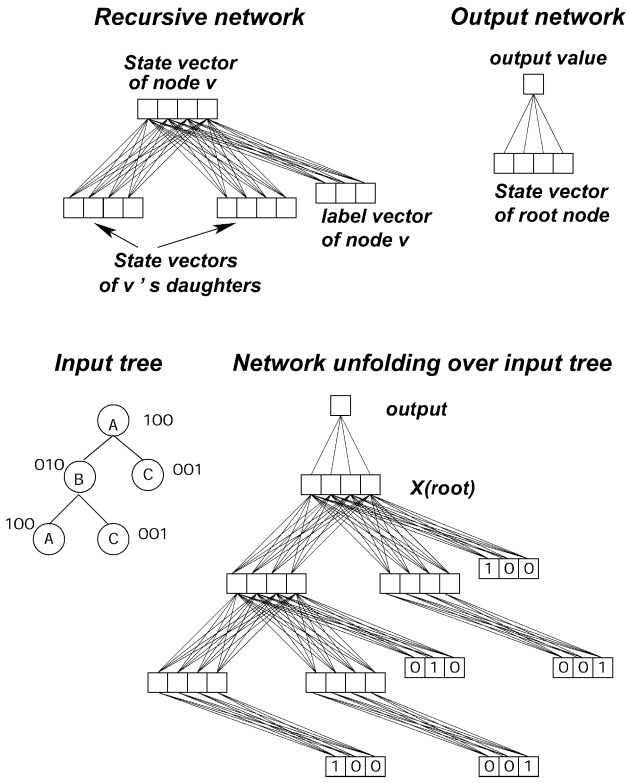

Fig. 5. The network architecture. The top panel shows the recursive and output networks. The bottom panel shows the recursive network "unfolded" according to the topology of a simple input tree.

(vector-encoded) label[2] of v with the outputs of the recursive networks of each of v's daughters.[3] The weights on the connections are the same at each replication of the network in the unfolding. Note that the label vector and hidden state vector play similar roles to the input layer and context layer of simple recurrent networks (SRNs) (Cleeremans, Servan-Schreiber, & McClelland, 1989; Elman, 1990). In an SRN, the input layer encodes the label of a new input symbol S in the left-to-right processing of a sequence, and the context layer encodes the representation of the substring preceding S. The SRN combines these two sources of information to create the context layer for the next input symbol. In the case of RNNs, the label vector encodes the label of a new node N in a bottom-up traversal of a tree, and the RNN combines this information with the hidden state vectors of each of N's daughters. The result is a new hidden state vector which

[2] The labels are encoded with localist "one-hot encoding". This means that each label (corresponding to a part-of-speech tag or syntactic category) is encoded as a unique vector, whose components are all zero except for a single 1. The position occupied by the 1 in the list uniquely identifies the category.

[3] The recursive network incorporates positions for a fixed maximum number of daughters. In the experiments reported here, we allow a maximum of 15 daughters for any given node.

encodes the subtree dominated by *N*. Thus, RNNs can be seen as a generalization of SRNs, where one is dealing with trees instead of just sequences.[4] In their use of recursive distributed representations, RNNs are similar to recursive auto-associative memories (RAAMs) (Pollack, 1990). RAAMs, however, can only encode and decode trees and cannot solve a supervised learning problem such as the one formulated here.

The *output network* is also a feedforward network, which implements a function taking as its input the hidden state vector $X(r)$ associated with the root r (see Fig. 5). The output of this function will differ according to the task for which the network is trained. A standard task for a RNN is tree classification (Frasconi et al., 1998), in which the network is trained to decide whether or not a given tree has some particular feature. In our experiments, we are evaluating a set of trees representing alternative attachments, rather than classifying a single tree, and this requires a change in the standard formulation of the learning task, as we discuss below.

We now consider the learning method and network architecture (a more detailed description can be found in Costa, Frasconi, Lombardo, & Soda, in press). In order to create a training set, we take a sample of sentences from the treebank, and, for each word w_i, we take the previous incremental tree T_{i-1} and find all the possible attachments of w_i to T_i that can be made using the connection paths in the grammar. This results in a set of candidates for T_i. We call this set of alternatives for attaching w_i the *forest* F_i. For example, if the training set includes a sentence with a relative clause ambiguity such as *The servant of the actress who…*, the forest for the word *who* will include trees representing all possible attachments of this word, including the attachments to the high and low noun phrase sites. One of these attachments will correspond to the correct attachment for the word *who* in the treebank corpus, and the incremental tree representing this attachment will be the learning target. The trees representing the other attachments will be used as the negative examples in training.

More formally, each training example is a pair (F_i, j^*), where F_i is the forest of alternatives, and j^* is the index of the correct tree (T_i) in F_i. The RNN is trained to predict the conditional probability that a tree T_i is the correct one, given the entire forest F_i:

$$y_{ij} = P(T_j - T_{j^*} | F_i) \text{ for each } j \in \{1, \ldots, \|F\|\}$$

where j is the index of a tree in the forest, j^* is the index of the correct tree, and y_{ij} is the probability estimated for the tree indexed j in the forest F_i. Assuming the correct tree belongs in the forest F_i (i.e. assuming the grammar induced from the database of all available connection paths is complete), the probabilities for all of the candidate trees must sum to 1, i.e. the following holds:

$$\sum_j y_{ij} = 1$$

Fig. 6 summarizes the stages involved in calculating the probability estimates for a forest of alternative attachments. We use an RNN to process each tree in the forest of

[4] In fact, RNNs can process directed acyclic graphs, of which trees are a sub-class (Frasconi et al., 1998; Goller & Kuechler, 1996; Sperduti & Starita, 1997).

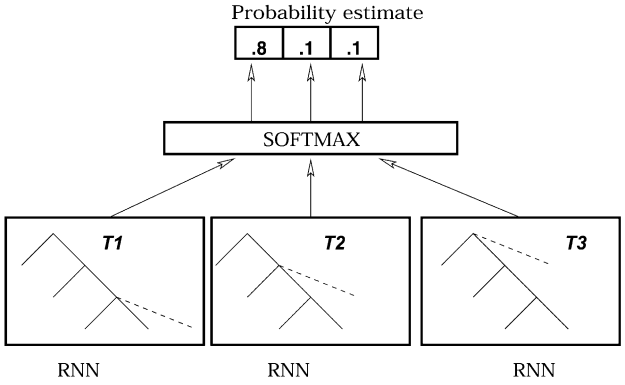

Fig. 6. Calculating probability estimates for a forest of three alternative attachments. In this example, a connection path (illustrated with the dashed line) can be attached to three different anchors in the preceding incremental tree, yielding three different possible incremental trees (T1, T2, T3).

alternative attachments for each word in the treebank, maintaining the same weights in the networks for each tree. Moreover, the output function applied to each RNN is linear, yielding a real number a_{ij} associated with the j-th tree in F_i. All the linear outputs are finally transformed using the softmax function (normalized exponentials), yielding the estimates of the conditional probabilities y_{ij} for each tree in the candidate forest. Normalizing in this way forces the candidate trees in the forest to compete against each other, so that the final probability estimate for some tree T will depend not only on the configurational features of T itself, but also on the scores given to other trees in the forest.

During training, in a feedforward phase, information flows bottom up through each tree, and the normalized output of each tree is compared with a learning target (1 if the tree corresponds to the actual tree T_{j*} in the treebank, and 0 if it does not).[5] The errors are calculated from the output of each tree in relation to the learning target, and are then back-propagated through the RNN representing each corresponding tree, using the back-propagation through structure algorithm (Goller & Kuechler, 1996; Sperduti & Starita, 1997). Training maximizes the conditional log-likelihood of the predicted preferences, given the true tree T_{j*}. The network is exposed to the training set for repeated epochs until performance on a validation set (unseen during training) is maximized.

After training, the network can be used in a purely feedforward fashion to give a first-pass attachment preference, by estimating the probabilities for each tree in any given forest (as in Fig. 6).

5. Experiments

We evaluated the performance of the model both on unrestricted text and on psycho-

[5] Here we assume that T_{j*} can be obtained from a treebank, using the procedure described in Lombardo and Sturt (2002).

linguistic test sentences. For the experiments, we trained the network using the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993). We used the method described in Lombardo and Sturt (2002) to collect a grammar of 1675 connection paths from a sample of 2000 sentences randomly selected from the treebank. The network was then trained on 500 of these sentences. The training set was created by taking each word w_i in the treebank, with the previous correct incremental tree T_{i-1}, and computing the forest of alternatives for that word F_i, using the grammar of connection paths. Each forest contained exactly one correct incremental tree T_i. The network generated for each epoch of training was validated on 100 unseen sentences, and the validation yielded a single network which achieved the best generalization performance. This network was then tested on another 500 sentences from the treebank (unrestricted text). Note that both the test set and the training set were drawn from the 2000 sentences used to create the set of connection paths. This means that the correct connection path could be guaranteed to be included in at least one of the alternative incremental trees in each forest that was considered in the test set. The model was also tested on a collection of psycholinguistic examples, drawn from the experimental sentence-processing literature. The psycholinguistic sentences were not included in the 2000 sentences used to create the set of connection paths. Note that the task of the network is to make a first-pass attachment decision for the part-of-speech category of each word w_i, given the correct incremental tree T_{i-1}. The model is not a full parser, in that it does not include any mechanism for storing or recovering alternative analyses. A full parser based on the network is, however, currently under development. Recall also that we are working with trees whose terminal nodes are part-of-speech categories, rather than actual words; thus, the model is lexically blind. The trees in the corpus used for training, validation and testing were transformed by "flattening" all local trees in which left recursion occurred. For example, local trees like 1 were transformed into local flatter trees like 2:

(1) $[_{NP}\ [_{NP}$ DT NN] PP]
(2) $[_{NP}$ DT NN PP]

The reason for this is that, in our incremental processing model, left-recursive structures are processed using an adjunction-like operation, in which a connection path is "inserted" into the right frontier of an incremental tree, rather than being attached in the standard sense. This requires splitting an anchor into two in the incremental tree, and the model is not currently configured for such situations.[6] For the purposes of testing, the network was given the part-of-speech category for each word w in the test sentence, with the corresponding correct incremental tree T_{i-1}. The task was to rank the forest of alternatives generated by the grammar. The highest ranked T_i was checked against the actual T_i in the test sentence (for the unrestricted text), or against the T_i that corresponds to the human first-pass preference (for the psycholinguistic examples).

[6] However, there is no theoretical reason why our architecture could not deal with this insertion operation.

5.1. Experiment 1: unrestricted text

The unrestricted text consists of the 500 sentences in the training set. For the unrestricted text, the average number of candidate trees for attaching a word to the incremental tree was 56, and the maximum was 600. Clearly, picking the best candidate tree in the face of such pervasive ambiguity is a non-trivial task. The network's performance is compared with two baseline conditions in which the ranking was computed using the well-known first-pass attachment principles Minimal Attachment and Late Closure (Frazier, 1978; Frazier & Rayner, 1982; Kimball, 1973). Minimal Attachment is a preference for simple structures, and in terms of our model is computed by finding the shortest connection path. Late Closure is a preference for attachments to recently processed material, and in our model corresponds to a preference for an anchor which is as low as possible in the tree. The baseline conditions consisted of an MA-LC condition, and a LC-MA condition. These conditions corresponded to different heuristics that would be applied in the case of a tie. In the MA-LC condition, Minimal Attachment took priority over Late Closure in the case of a tie. In other words, when Minimal Attachment and Late Closure each expressed a preference for a different attachment, the Minimal Attachment preference was taken. The LC-MA condition was the converse of this, with Late Closure taking priority in the case of a tie. Note that all of these preferences were computed on the incremental trees, rather than the final trees. It can also be noted here that the MA-LC condition corresponds to Frazier's claim that Minimal Attachment is prioritized over Late Closure (Frazier, 1987).

The results for the unrestricted text are summarized in Fig. 7. The graph shows the percentage of times that, after sorting the candidates in the forest, the correct tree is found within the first k trees, where k ranges between 1 and 15. It can be clearly seen that the network performs consistently better than the two baseline conditions. Perhaps this is not surprising; after all, people do occasionally produce sentences that include non-minimal, and non-recent attachments, and we would expect an experience-based model to be able to

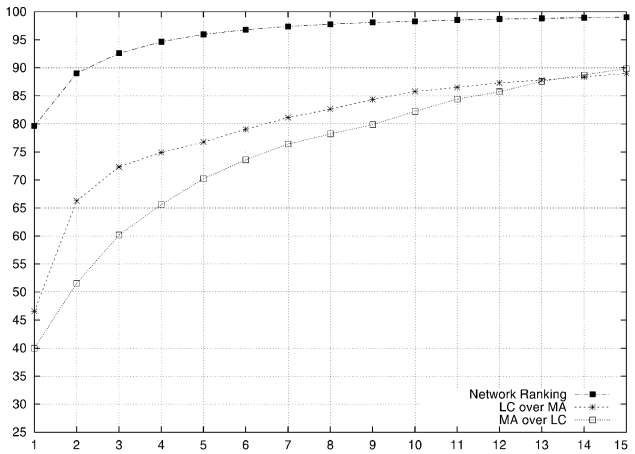

Fig. 7. Results for unrestricted text. The graph shows the percentage of times that, after sorting the candidates, the correct tree is found within the first k trees, for $k = 1, \ldots, 15$. The network's ranking is compared with heuristics based on Late Closure and Minimal Attachment (see text for details).

learn such contingencies. However, it is certainly true that simpler attachments are in general more frequent than complex attachments, and it is probably true that recent attachments are in general more frequent than non-recent attachments. A frequency-based model which worked at a particularly coarse-grained level might therefore learn only to apply these two heuristics, raising the possibility of producing a network that is empirically indistinguishable from Garden Path theory. The results of this experiment show clearly that this is not the case for the RNN, showing that it is worthwhile to investigate it further as an experience-based model.

Incidentally, it can be seen that the LC-MA condition yielded better performance than the MA-LC condition. This appears to militate against the claim by Frazier (1987) that Minimal Attachment is prioritized over Late Closure in human first-pass attachment. However, we should not make too much of this pattern of results; recall that the treebank from which we extracted the training and test set was "flattened" to remove left recursion. This flattening has the effect of converting more complex structures into simpler structures. The flattening procedure may therefore have artificially created many situations in which two alternative attachments were equally simple, but where in the original treebank they were not. In these situations, Minimal Attachment would have been applicable in the original treebank example, but not in the example transformed by the flattening procedure.

The results of the network are extremely encouraging, as the correct tree is ranked first in the candidate list (position 0) around 80% of the time, and in the first three positions 93% of the time.

Further comparisons were made to evaluate the ability of the network to generalize to unseen input. To do this, we calculated the number of trees in the test set that had appeared in the training set. We did two comparisons. The first calculation considered only the correct incremental trees (i.e. those incremental trees that correspond to the actual attachments in the treebank), and the results are shown in the table below:

Trees in test set	11,011
Trees in training set	11,250
Trees from test set in training set	420
Percentage	4%

A second calculation combined the trees from all the forests that had been seen in training, and all the forests that were encountered in the test set. Thus, in addition to the correct incremental trees, this comparison also included negative examples.

Trees in test set	480,928
Trees in training set	517,308
Trees from test set in training set	4469
Percentage	1%

The small percentage of the training set seen in the test set clearly illustrates the sparse data problem. However, the network's performance shows that the model overcomes the

problem by generalizing from past experience to unseen examples. Clearly, the model's performance is superior to any naive method of estimating probabilities based solely on experienced examples of actual trees. We will demonstrate below (Experiment 3) that the network exhibits systematic preferences in the face of unseen ambiguities.

These results clearly show that the network is a valid informer for incremental parsing decisions. Of course, to work effectively the parser needs to be equipped with a recovery mechanism in order to proceed appropriately when the first-pass selection is wrong and the error is recoverable.

5.2. Experiment 2: psycholinguistic preferences

In this section, we will present an evaluation of the model's performance on psycholinguistic examples. The intention is to look at cases in which structural preferences have been argued to apply on the basis of controlled experiments, and to check the network's ranking of candidate incremental trees for the critical words that represent ambiguous attachments. We then compare the ranking of incremental trees corresponding to the preferred and dispreferred alternatives. Given the difficulty of disentangling structural factors from lexical and other non-structural influences on disambiguation, it might be argued that some of the ambiguities mentioned below are not really cases where structural preferences apply. However, we believe that all the ambiguities mentioned below are cases in which the Tuning Hypothesis would argue for a first-pass preference based on structural frequencies, and it would be reasonable to assume that in a multiple constraints model such as that of McRae et al. (1998), a structural frequency-based preference would be one of the constraints influencing the resolution of the ambiguity. One fact that will be noticed is that many of the examples have more possible readings at the onset of the ambiguity than is standardly assumed in psycholinguistic research. Thus, the procedure gives an accurate estimate of the degree of ambiguity that should be expected from a lexically-blind, but realistically wide-coverage system. Note also that since the model deals with purely structural non-lexical ambiguities, we do not include part-of-speech ambiguities, such as relative pronoun vs. complementizer uses of the word *that*, or past participle vs. past tense uses of verbs.

5.2.1. Replication and statistical significance

For ease of exposition, we only give detailed discussion for the results of the single network trained on the 500 sentences as discussed above (we will refer to this as the *base network*, and its training set as the *base training set*). However, we have also used the same training procedure to produce 20 other networks (which we will call *replication networks*), each of which was trained on a different random sample of the Wall Street Journal, Penn Treebank, and each of which was initialized with a different set of random weights before training began. The purpose of this procedure was to rule out the possibility that the preferences expressed by the network are the result of arbitrary and unknown features of the 500 particular sentences selected for training, or that they arise from the particular configuration of random weights used to initiate training. For each of the ambiguities discussed, Wilcoxon's signed ranks test was used to determine whether the preference

for one or other reading of the ambiguity was systematic across the 20 replication networks.[7] The results of these statistical tests will be given for each of the ambiguities.

5.2.2. Frequency in the Treebank

In discussing the results for the psycholinguistic preferences, it is interesting to consider the relation between the network's preferences and the actual frequency of constructions in the training set. To shed light on this issue, we will also report searches that have been run on the training set, to calculate the relative frequencies of the configurations corresponding to the different readings of each ambiguity. The results of these searches will be given for each ambiguity. In defining the search configurations, we had to confront the grain problem – how detailed should the search configurations be? In most cases, the exact incremental tree to which the network is asked to make its attachment will not have occurred at all in the training set. Therefore, we attempted to search for the portion of the incremental tree that seems relevant to the ambiguity being studied, defining the search configurations in terms of right frontiers and incremental trees. Specifically, to test the relative frequencies of two possible attachments for some word w_i, we take the right frontier of the previous incremental tree T_{i-1}, from the lowest attachment site to the highest attachment site being compared, plus the connection path. In some cases, where the search returned only a very small number of examples, the search was re-run with a less detailed configuration to gain a larger sample, and thus a more realistic idea of the frequencies in the training corpus. We will give the definitions of search configurations for each ambiguity.

5.2.3. NP/S ambiguities

Our first example will be concerned with the local ambiguity in which a noun phrase can be interpreted either as the object of a preceding verb, or as the subject of a complement clause. An example from Pickering, Traxler, and Crocker (2000) is given below:

4. The athlete realized his goals were out of reach.

This ambiguity has been very extensively studied (Frazier & Rayner, 1982; Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Pickering et al., 2000; Rayner & Frazier, 1987; Trueswell, Tanenhaus, & Kello, 1993). Initial studies (Frazier & Rayner, 1982; Rayner & Frazier, 1987) showed convincing evidence for an initial preference for the object reading of the noun phrase, and although this preference can be modulated by subcategorization biases of the verb (Garnsey et al., 1997; Trueswell et al., 1993), by plausibility (Pickering & Traxler, 1998), and by an interaction of the two (Garnsey et al., 1997), there is very little evidence suggesting that the preference can actually be reversed. Furthermore Pickering et al. (2000) and Kennison (2001) show that a preference for the object reading can be found even when the subcategorization of the preceding verb is biased towards the alternative

[7] Wilcoxon's signed ranks test was used because it does not depend on any assumptions about the statistical distribution of the data (in our case, the probability estimates returned by the network). In all cases, Wilcoxon's test was applied with Yates' correction for continuity, and the significance level is expressed as a two-tailed probability approximation.

(though this is controversial; see Garnsey et al., 1997; Trueswell et al., 1993). Pickering et al.'s findings have been cited as support for the Tuning Hypothesis (Mitchell et al., 1995), since the direct object reading is more frequent, given a purely structural method of counting the corpus examples (see below).

We will now examine the network's treatment of this ambiguity. The following table gives information about the forest of incremental trees produced by applying the network's ranking for the attachment of (the part-of-speech category of) each word w_i, given the correct incremental tree T_{i-1}. For each word, we show the position of the "correct tree" in the ranking, and the number of trees in the forest. Note that the "correct tree" is the one that is consistent with the overall global reading of the sentence (in this case, the complement clause reading, which does not correspond to the initially preferred direct object reading). The critical word, where the attachment decision of interest has to be made on an incremental model, is marked in bold.

	The	athlete	realized	**his**	goals	were	within	reach.
Position of "correct" T_i	1	1	1	**2**	1	1	1	1
Cardinality of Forest	1	1	3	**5**	32	15	90	43

At the critical word, the network had to decide between five trees generated by the attachment of that word to the previous incremental tree. Of these five, two are of interest for the current investigation, and they are shown in Fig. 8, which also shows the search configuration used for testing the frequency of these two constructions in the training set (see below). From the table, it can be seen that the ranking of the globally correct tree (i.e. the tree corresponding to the human-dispreferred complement clause reading) is second on a list of five alternatives. The tree which is ranked first, however, corresponds to the noun phrase object reading, which is preferred by humans in the first-pass analysis, and is the more frequent, as we have seen above.[8] We can gain a more detailed impression of the preference of the network by looking at the actual probability estimates returned by the normalized output of the network. For the direct object reading and clause reading at the critical word, the estimates were as follows:

	Probability estimate	Ranking
Direct object reading	0.93	1st
Clause reading	0.06	2nd

Thus, the network makes the correct prediction here. Incidentally, it will be noticed that the position of the correct tree is consistently high on the list of alternatives throughout the sentence. The globally correct tree is at the top of the list except for the critical word, where humans prefer the alternative direct object analysis.

Statistical analysis compared the probability estimates of the direct object and sentential complement readings for each of the 20 replication networks (see above) using a Wilcoxon

[8] Remember that all input words in the simulation are given the left context corresponding to the globally "correct" parse. That is why the sentence complement analysis becomes the highest ranked alternative at "goals".

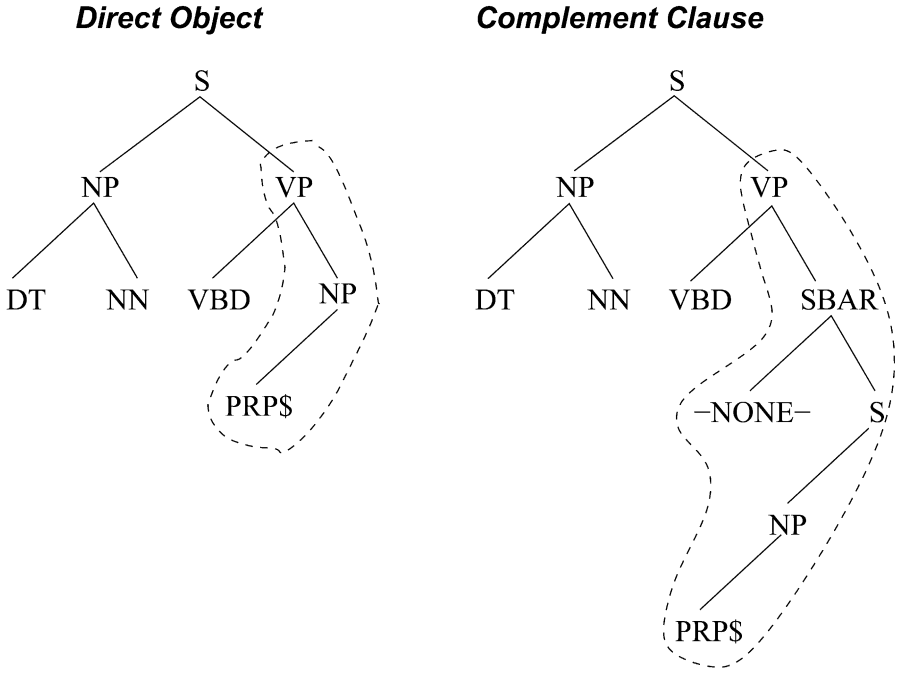

Fig. 8. Incremental trees representing the direct object and complement clause readings for the NP/S ambiguity. Treebank conventions are: PRP$ = possessive pronoun (e.g. *his*), VBD = past tense verb, -NONE- = empty category (in this case an empty complementizer). The dotted loops enclose the configurations that were used for searching the frequencies of these two constructions (see text).

signed ranks test for paired samples. This showed that the difference between the two attachments was statistically reliable across the 20 networks (direct object mean: 0.64, sentential complement mean: 0.33, $T(N = 20) = 39$, $P < 0.05$). Thus, the preference is robust across different training sets and random weight configurations.

To test the relative frequencies of the two relevant constructions in the training set, we searched for examples of the two configurations enclosed in the dotted loop in Fig. 8. The results were that, in the 500-word base training set, there were 40 examples of the direct object reading and three examples of the complement clause reading. The frequencies for the combined training sets of the replication networks replicated this difference (direct object: 675, complement clause: 62). Thus, the preference expressed by the network clearly mirrors the frequencies in the training set for this ambiguity.

5.2.4. Relative clause attachment

Phenomena involving the attachment of relative clauses were the original motivation for the Tuning Hypothesis. As we have mentioned above, Cuetos and Mitchell (1988) found that, where a relative clause could modify either of two noun phrases, English speakers showed a low attachment preference, preferring to modify the noun phrase whose head had most recently been read. That is, in (5), the preference will be for an

interpretation in which the actress, rather than the servant, is on the balcony. Note, however, that although this preference is often discussed with relation to the Tuning Hypothesis, the strength of the preference appears to vary according to various factors, such as the type of the preposition used in the complex noun phrase (Gilboy, Sopena, Clifton, & Frazier, 1995; Traxler et al., 1998). Since the model described here uses purely non-lexical information, such distinctions cannot be considered.

5. The servant of the actress who was on the balcony died.

Below we give the information about the position of the correct tree in the candidate set for each word and its preceding incremental tree.

	The	servant	of	the	actress	**who**	was	on	the	balcony	died.
Position of "correct" T_i	1	1	1	1	1	**1**	1	1	1	1	1
Cardinality of Forest	1	1	20	25	18	**4**	7	89	76	50	16

For the critical word, the two relevant candidate trees corresponding to the high and low attachments are illustrated in Fig. 9, which also shows the configurations used to search for the relative frequencies of these constructions in the training set (see below). Again, looking at the results for each incremental tree, it can be seen that for all words, the correct incremental tree is ranked at the top of the list of candidates, including the preferred low attachment of the critical word (marked in bold).

The probability estimates and rankings of the high and low attachments at the critical word *who* are summarized below:

	Probability estimate	Ranking
Low attachment	0.70	1st
High attachment	0.30	2nd

Again, the 20 replication networks repeated this pattern, with a mean probability estimate of 0.82 for the low-attached reading and 0.18 for the high-attached reading ($T(N = 20) = 5, P < 0.05$), showing that the preference was not due to any idiosyncratic biases in the training set.

Using the search configurations illustrated in the dotted loop in Fig. 9, we found again that this preference was reflected in the corpus frequencies. In the 500-word base training set, there were eight examples of high attachment and 21 examples of low attachment. For the combined training sets of the 20 replication networks, there were 143 examples of high attachment and 351 examples of low attachment.

5.2.5. Prepositional phrase attachment to noun phrases

Traxler et al. (1998) studied attachment ambiguities which are similar to the relative clause example given above, but where the constituent to be attached is a prepositional phrase rather than a relative clause. They found a low attachment preference for this ambiguity.

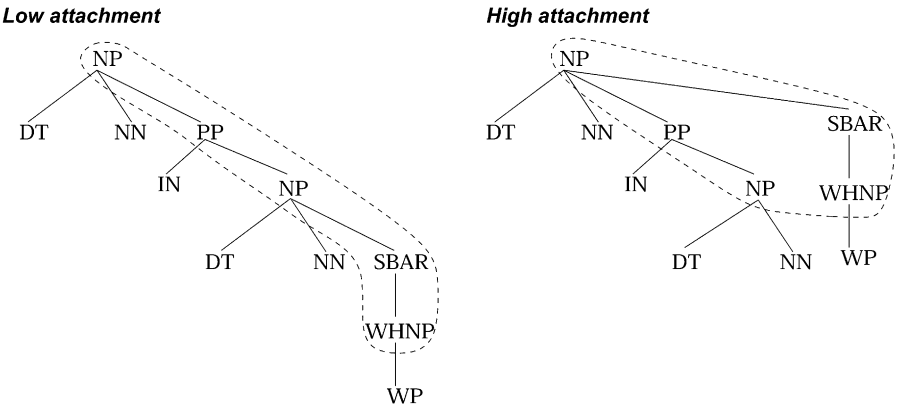

Fig. 9. Incremental trees representing the low and high attachments for the two-site relative clause ambiguity. The dotted loops enclose the configurations that were used for searching the frequencies of these two constructions (see text).

This preference is also found in the network's ranking:

	The	servant	of	the	actress	**with**	the	wooden	leg	died.
Position of "correct" T_i	1	1	1	1	1	**1**	1	1	1	1
Cardinality of Forest	1	1	20	25	18	**41**	47	45	28	7

The incremental trees representing high and low attachment were analogous to those for the two-site relative clause ambiguity discussed above. Here, it can be seen that the correct incremental tree remains at the top of the list of alternatives for each word processed, and, in particular, the correct low attachment preference for the word *with* is reproduced.

The probability estimates and rankings for the low and high attachments at the critical word *with* are given below:

	Probability estimate	Ranking
Low attachment	0.73	1st
High attachment	0.13	2nd

Again, the replication networks also showed this pattern, with a mean probability estimate of 0.85 for the low-attached reading and 0.1 for the high-attached reading ($T(N = 20) = 1$, $P < 0.05$).

As in the ambiguities discussed above, the low preference generally reflected the frequencies in the training set; using search configurations analogous to those defined for the two-site relative clause ambiguity discussed above, it was found that, in the 500-word base training set, there were 25 high-attached pps and 73 low-attached pps. In the combined training sets for the replication networks, there were 580 high-attached, and 1544 low-attached examples.

5.2.6. Adverb attachment

There has long been an intuition that adverbs are preferentially attached low (Kimball, 1973). For example, in the following sentence, the preference is for the adverb *miserably* to modify *failed* rather than *said*:

6. John said that he failed miserably.

In a series of eyetracking experiments, Altmann, van Nice, Garnham, and Henstra (1998) confirmed this preference, although they also showed that the preference can be overridden given very strong contextual manipulations.

Results for the candidate sets of this sentence are given below:

	John	said	that	he	failed	**miserably**.
Position of "correct" T_i	1	2	1	1	1	**2**
Cardinality of Forest	1	3	42	10	15	**122**

The two critical incremental trees are illustrated in Fig. 10. It can be seen that the preferred low attachment of the adverb is ranked second out of a list of 122. The first ranked tree for this word differed from the actual preferred tree only in the connection path, and not in the anchor. The difference was minor, and related to the maximal projection of the adverb. The actual correct connection path (ranked second) included an adverbial phrase (ADVP) as a maximal projection. In the connection path that was ranked first by the network, there was no maximal projection, but the adverb was still attached low. By

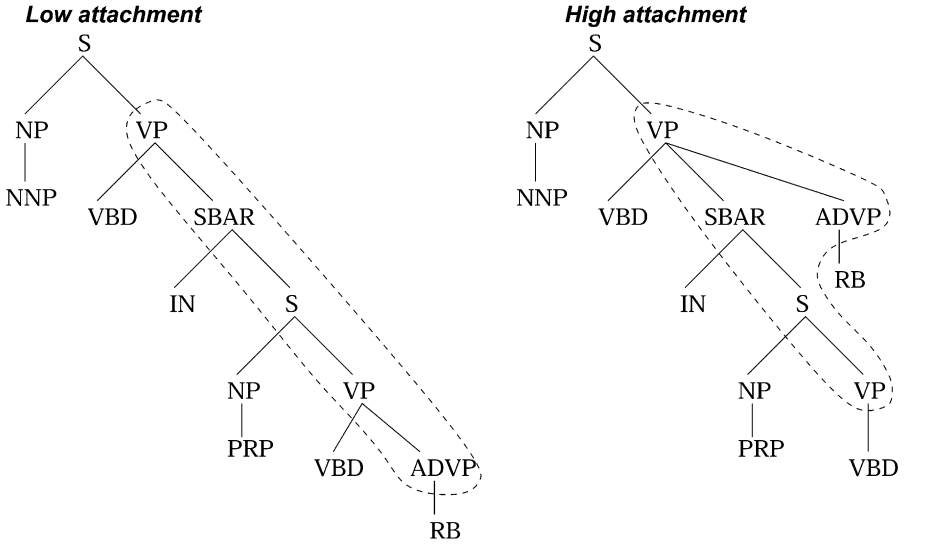

Fig. 10. Incremental trees representing the low and high attachments for the adverb attachment ambiguity. The dotted loops enclose the configurations that were used for searching the frequencies of these two constructions (see text).

comparison, the alternative high attachment of the adverb was ranked 20th out of the 122 alternatives, as the following shows:

	Probability estimate	Ranking
Low attachment	0.23	2nd
High attachment	< 0.01	20th

The first ranked tree, in which the adverb was attached directly to the lower verb phrase without an intervening maximal projection, was assigned a probability of 0.47 by the model.

Statistical analysis on the replication networks again replicated this pattern, with a mean probability estimate of 0.29 for the low-attached alternative, and an estimate of below 0.01 for the high-attached alternative ($T(N = 20) = 0$, $P < 0.05$).

Again, the low attachment preference was reflected in frequencies of the training set. Based on the search configurations shown in the dotted loop in Fig. 10, the search revealed that, in the 500-sentence base training set, there were two examples of low attachment and 14 examples of high attachment. In the larger sample of the combined replication training sets, there were 30 examples of high attachment and 350 examples of low attachment.

5.2.7. Closure ambiguity

There is a lot of evidence suggesting that, in sentences like (7), people prefer to attach the noun phrase *the sock* as the object of the preceding verb, rather than as the subject of the main clause (Frazier & Rayner, 1982):

7. While Mary knitted the sock it fell.

Moreover, Mitchell (1989) presented evidence suggesting that this preference exists even when the preceding verb is obligatorily intransitive. Therefore, this is a good candidate for a structural preference.

There was one occurrence of this ambiguity in the 500-sentence training set (where the relevant noun phrase was attached low). In the combined training sets for the replication networks, there were 29 occurrences of low attachment and no occurrences of high attachment (Fig. 11). Results for the candidate sets of this sentence are given below:

	While	Mary	knitted	**the**	sock	it	fell.
Position of "correct" T_i	1	1	1	**1**	1	4	1
Cardinality of Forest	6	1	7	**30**	30	19	6

Again, it can be seen that the preferred reading of the ambiguity is ranked at the top of the list for the critical word *the*, in this case, as the first of 30 alternatives. Probability estimates and rankings are given below:

	Probability estimate	Ranking
Low attachment	0.97	1st
High attachment	< 0.01	24th

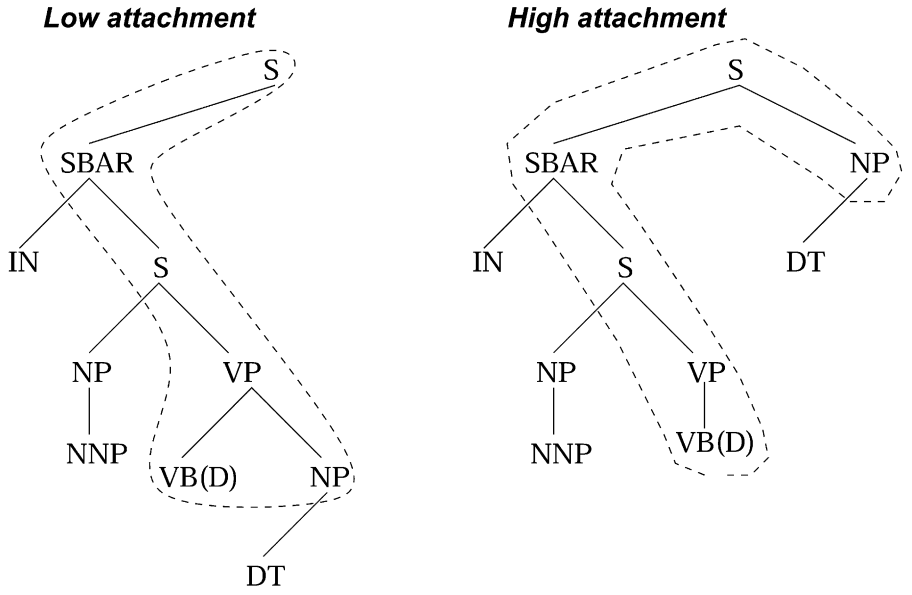

Fig. 11. Incremental trees representing the low and high attachments for the closure ambiguity. The dotted loops enclose the configurations that were used for searching the frequencies of these two constructions (see text). The symbol VB(D) means that a past tense verb (VBD) was included in the configuration used to test the network, but any verb (VB*) was allowed in the treebank search configuration (see text).

The pattern was again replicated in the statistical analysis of the replication networks, with a mean probability estimate of 0.78 for the low-attached reading, and 0.01 for the high-attached reading ($T(N = 20) = 0$, $P < 0.05$).

Searches for frequency of occurrence showed that the low attachment reading was more frequent, and the high attachment reading occurred only extremely rarely. In defining the search configuration for this ambiguity, it is important that the ambiguous noun phrase (corresponding to *the sock* in the above example) should immediately follow the embedded verb *knitted*, since the crucial point is that this is an ambiguity between a transitive and an intransitive use of the verb. For reasons of sparse data, the search configuration was defined so that it matched any verb, while the network was tested on a configuration which specified that the verb was past tense[9] (see Fig. 11).

In the 500-sentence training set, there were seven occurrences of low attachment, and no occurrences of high attachment. In the combined training sets for the replication networks, there were 212 examples of low attachment and two examples of high attachment. Thus, the strong low attachment bias mirrors the distribution in the training sets.

5.2.8. PP attachment ambiguities

Ambiguities in which a prepositional phrase can be attached to either a noun phrase or a verb phrase have received much discussion in the literature. Rayner, Carlson, and Frazier

[9] Verbs in the Penn Treebank are labelled with tense information as part of the atomic part-of-speech category.

(1983) examined sentences such as the following:

8. The spy saw the cop with the binoculars.

In their eyetracking experiments, Rayner et al. (1983) found evidence for an initial preference for the verb phrase attachment (i.e. the interpretation in which it is the spy rather than the cop who has the binoculars). Since that time, a number of studies have found that this preference can be modulated by non-syntactic factors such as the type of verb involved (Konieczny, Hemforth, Scheepers, & Strube, 1997; Spivey-Knowlton & Sedivy, 1995), discourse felicity (Altmann & Steedman, 1988) and the argument status of the prepositional phrase (Schuetze & Gibson, 1999).[10] In the Altmann and Steedman (1988) study, the overall preference was actually for the noun phrase attachment, contra to Rayner et al.'s results. Interestingly, in corpora, the verb phrase attachment is actually less frequent than the noun phrase attachment, when a method of counting is employed which ignores lexical items (Hindle & Rooth, 1993). This difference was also found in our training set (see below).

Thus, it was expected that the network would reproduce this noun phrase attachment reading. As can be seen from the results at the word *with*, this preference was found in the network, though it is interesting to note that the probability estimates are fairly close.

	The	cop	saw	the	spy	**with**	the	binoculars.
Position of "correct" T_i	1	1	1	3	1	**2**	1	1
Cardinality of Forest	1	1	3	22	25	**57**	29	28

	Probability estimate	Ranking
NP attachment	0.49	1st
VP attachment	0.39	2nd

The statistical analysis for the replication networks showed that the NP attachment preference is reliable (mean for NP attachment 0.64, and VP attachment 0.26, $T(N = 20) = 24$, $P < 0.05$).

In order to determine the actual frequencies in the training set, we used the search configurations illustrated in Fig. 12. For the 500-sentence main training set, there were 98 instances of noun phrase attachment and 34 instances of verb phrase attachment. For the combined replication training sets, there were 1619 examples of noun phrase attachment and 623 instances of verb phrase attachment.

Interestingly, when the search criteria were slightly changed, this pattern reversed. In the revised search criteria, the configuration did not include the noun part-of-speech category (marked NN in Fig. 12). Thus, the head daughter of the noun phrase was not specified. With this revised definition, there were 169 examples of NP attachment in the 500-sentence base training set, and 183 examples of VP attachment, and this pattern was

[10] Note, however, that we cannot consider these lexical factors in the model in its current form, because we do not include lexical information in the representations.

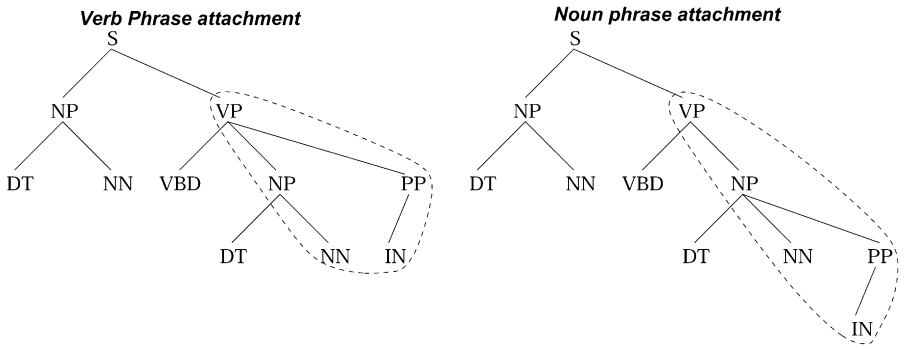

Fig. 12. Incremental trees and search configurations for the PP attachment ambiguity.

also found in the replication training sets (NP: 2812 pp-np, VP: 3248). This striking difference illustrates the dangers of using arbitrarily-chosen pre-defined configurations as a basis for counting structural frequencies, and motivates the use of models such as the one proposed here, which do not rely on pre-defined search configurations.

5.2.9. Three-site relative clause ambiguities

Three-site relative clause ambiguities such as (9) represent a challenge for experience-based theories of ambiguity resolution.

9. The friend of the actress of the servant who was on the balcony died.

In such ambiguities, the relative clause could be attached to one of three possible sites, the low site (the servant was on the balcony), the middle site (the actress was on the balcony), and the high site (the friend was on the balcony). It has been convincingly shown by Gibson, Pearlmutter, Canesco-Gonzalez, and Hickok (1996) that the human preference ordering for this ambiguity is low–high–middle, while available corpus data appear to show a preference ordering of low–middle–high, although the rare occurrences of the relevant configurations in existing treebank corpora make it difficult to judge (see below).

	The	friend	of	the	servant	of	the	actress	**who**	was	on	the	balcony	died.
Position of "correct" T_i	1	1	1	1	1	1	1	1	**1**	1	4	1	1	1
Cardinality of Forest	1	1	20	25	18	41	47	28	**7**	9	110	98	60	18

The following table shows the network's estimation for the ambiguity.

	Probability estimate	Ranking
Low attachment	0.71	1st
Mid attachment	0.05	3rd
High attachment	0.24	2nd

It can be seen that the model trained on the base training set predicts the correct human

low attachment preference for this ambiguity. In addition, the ranking shows an ordering corresponding to the human preference ordering.

The analysis of the preferences in the replication networks showed a similar numerical pattern (mean low: 0.75, mean middle: 0.10, mean high: 0.14), but the relevant differences were not all statistically reliable. The differences between low and middle, and between low and high attachments were both reliable (both Ts < 8, both Ps < 0.05). However, the difference between high and middle attachments was not ($T(N = 20) = 68$, $P > 0.1$). Thus, although it is safe to conclude that the network has a reliable preference for low attachment as its first choice, we cannot make any firm conclusions about the relative preferences for the network's second and third choices.[11]

We ran searches using search configurations analogous to those used for the two-site relative clause ambiguities (see Fig. 9), to examine the relation between the preference expressed by the model and the actual frequencies in the corpus. In the 500-sentence base training set, we found only one example of a high-attached three-site relative clause, and no examples of either middle- or low-attached relative clauses. Thus, at this level of granularity, three-site relative clauses are extremely rare. However, it is possible that the base network's preference for high over middle attachment was caused directly by its experience of the single sentence exhibiting a high-attached relative clause ambiguity with three sites. To evaluate this, we took this one high attachment sentence, and re-attached the relative clause to the middle attachment site in the training set. The other sentences in the training set remained exactly as in the original training. A new network was then trained on this altered training set, and this network also exhibited a low–high–middle attachment ordering (probability estimates; low: 0.60, mid: 0.10, high: 0.28). Thus, the base network's preference was not caused by the its experience of the single sentence in training which exhibited this ambiguity.

The larger sample of the combined training sets for the 20 replication networks gave a slightly more realistic picture of the frequencies; there were 42 examples of low attachment, 12 examples of middle attachment and seven examples of high attachment. This shows that the low attachment disambiguation is indeed more frequent in the corpus, but again, there is not really enough data for the high and middle attachments to make firm conclusions about frequencies.

To gain a better idea of the frequencies of the three-site relative clause attachments, we also counted frequencies with a much less detailed configuration, in which the critical noun phrase attachment sites did not have to be nested inside prepositional phrases. Specifically, we searched for sentences in which, during some point in processing, a relative clause could be attached to a previous incremental tree with three noun phrases on its right frontier. With this less detailed configuration, the counts were, for the 500-sentence base training set: high 1; mid 4; low 34, and for the combined replication training sets: high 9; mid 55; low 630. Thus, given a search definition that was sufficiently coarse-grained, the low–middle–high ordering was apparent.[12]

[11] It is possible that this lack of resolution is a consequence of the learning method; strictly speaking, the network is not trained to *rank* the candidate trees in the forest, but to pick out a single preferred tree.

5.3. Experiment 3: handling of sparse data

The above discussion of ambiguities from the psycholinguistic literature shows that the RNN architecture is able to form the basis of a viable model of experience-based structural preferences in ambiguity resolution. Of particular interest is the model's preference to attach new material to recently processed constituents in the tree. The ability to express this type of preference relies on being able to represent global information about the tree, and the relative positions of alternative attachment sites in it. Recall that it is precisely this type of preference which is difficult to express in formalisms that rely on more local information, such as probabilistic context-free grammars.

It should be noted that nearly all of the preferences of the model appear to be transparently related to the frequency of the relevant constructions in the training set. Thus, the model appears to implement a fairly direct frequency-based notion of learning from experience. It is therefore important to ask whether any characteristic behaviours emerge from the model which can differentiate it from other possible frequency-based models, assuming that such models would also be capable of learning the relevant configurational contingencies. In future work we intend to perform a number of systematic experiments on the model, to analyze such behaviours.

In this section, we describe an experiment which was designed to examine one of the characteristic behaviours of the model that we find particularly interesting, namely the way that the model generalizes over trees in order to handle ambiguities that have not directly been experienced during training. We have seen from the discussion of Experiment 1 that the test set of unrestricted text contained very few actual occurrences of trees that had been seen in the training set, yet the network was still able to make accurate predictions for attachment to those trees. Although humans obviously have vastly more syntactic experience than a 500-sentence training set, it is nevertheless reasonable to assume that humans often have to deal with syntactic configurations that are very rare, if not completely novel. Given this, it seems that the ability to generalize effectively must be an important property of any experience-based model of human disambiguation preferences.

In the experiment reported in this section, we will make a systematic study of the network's ability to handle unseen ambiguities, by directly manipulating the training set to which it is exposed. The experiment allows us to highlight one of the main differences between the model presented here and pure frequency-based accounts, such as those based on probabilistic context-free grammars.

5.3.1. Training set preparation

To demonstrate the network's handling of sparse data, we will concentrate on the two-site relative clause ambiguity discussed above. We took the original 500-sentence base training set, with its natural distribution of high and low relative clause attachments, and removed all sentences which exhibit a relative clause ambiguity, where a "relative clause

[12] Note that Gibson, Schuetze, and Salomon (1996) report extremely low frequencies for a search run on the entire Wall Street Journal corpus. However, their search configurations were much more detailed than ours, resulting in particularly sparse data.

ambiguity" is defined broadly as a sentence in which, at some point in processing, a relative clause could be attached to more than one possible noun phrase site on the previous right frontier. There were 45 such relative clause ambiguities.[13]

To create an adapted training set, all of these relative clause ambiguity sentences were replaced with new sentences, also from the Wall Street Journal Treebank, in which there was only one possible noun phrase to which the relative clause could be attached (for example, sentences like *The actress who was on the balcony died*). Thus, in this adapted training set, there were no two- or three-site relative clause ambiguities. A new network was trained on the adapted training set. The question was whether this new network would also exhibit a systematic preference for relative clause attachment, despite the fact that it had never been exposed to a relative clause ambiguity of the type considered here.

5.3.2. Testing the preference

In order to test whether any preference expressed by the network was systematic, we took a sample of 50 unseen two-site relative clause ambiguities from the Penn Treebank corpus, and in each case, asked the network to attach a relative pronoun to the previous incremental tree. For example, in the sentence *They met the director of the company that was in the news*, the network was given the incremental tree corresponding to the prefix *They met the director of the company*, and asked to attach the part-of-speech category of the relative pronoun *that*. Probability estimates for high and low attachments of each of the 50 relative clauses were analyzed using Wilcoxon's signed ranks test for paired samples, for both the original baseline network and the network trained on the altered data set. The mean probability estimates for each network were as follows:

	Baseline network	Sparse data network
Low attachment	0.64	0.67
High attachment	0.19	0.10

The low attachment preference was significant both for the baseline network and for the sparse data network (both $T\text{s}(N = 50) = 0$, both $P\text{s} < 0.05$). This demonstrates the ability of the network to generalize from its experience to deal with novel ambiguities.[14]

It is interesting to consider the possible mechanisms that might be behind the network's recency preference in the sparse data condition. One possibility is that, when information is lacking about the attachment of some phrase type, such as a relative clause, the network is able to generalize from its experience of the attachment of other phrase types. If these other phrase types in general show a low attachment preference, then the network will use this information to inform its attachment decision for unseen relative clause ambiguities. Preliminary results in recent experiments show that a RNN's experience of relative clause attachment has a small but reliable effect on its preferences for prepositional phrase

[13] Note that these include, but are not limited to the trees that match the configuration illustrated in Fig. 9.

[14] Recall that the means reported here are calculated from 50 test sentences, while the figures reported for Experiment 2 are the results for one test sentence. This is the reason for the small numerical difference in the degree of preference shown by the baseline network for this ambiguity in the two experiments.

attachment, thus demonstrating that this network architecture can indeed generalize across phrase types (Costa, Frasconi, Sturt, & Lombardo, 2002).

This ability to generalize to unseen ambiguities can be seen as a rather extreme case of what the network is doing in its normal operation; that is, generalizing from trees. We do not know how often humans have to deal with unseen ambiguities – in any case, whether an ambiguity should be classified as "unseen" depends on the granularity with which ambiguities are categorized. However, we do know that the individual incremental tree types encountered in processing realistic text are often extremely rare (see the discussion of Experiment 1). Thus, we believe it is highly likely that humans employ a mechanism that can generalize effectively at many levels.

The present experiment shows that the RNN model can generalize over trees in a theoretically interesting way. We believe that similar behaviour is a necessary component of any large-scale model of human sentence processing. This behaviour would not be expected, or would at least have to be stipulated, in any model based on the simple counting of configurations representing different ambiguities. The exact nature of syntactic generalization from experience in human syntactic processing is a topic that requires further experimental investigation, perhaps using a method such as syntactic priming in comprehension (Branigan, Pickering, Liversedge, Stewart, & Urbach, 1995).

6. General discussion

In this paper, we have presented an experience-based model of structural preferences in first-pass ambiguity resolution. The model relies on two key components. The first of these is a dynamic grammar, which draws a close correspondence between well-formedness rules on the one hand, and incremental parsing actions on the other. The second key component is a RNN model, which is a learning mechanism that deals directly with the natural representation of the data, i.e. syntactic trees. The network model performs tree generalization, therefore providing a solution to the problem of sparse data in the structural domain (see particularly Experiment 3). The use of the dynamic grammar allows us to represent attachment ambiguities in terms of the syntactic left contexts that are typically assumed in psycholinguistic theories of on-line ambiguity resolution. The result of this is that the network learns to discriminate between candidate incremental trees, which represent the partial structures that are computed during a left-to-right parse of the input. It may be that exposure to these intermediate structures allows the network to learn important structural generalizations that would be unavailable if it was exposed only to the full trees of the sentences in the corpus. In its treatment of the psycholinguistic examples, the model shows a fairly systematic preference for simplicity and recency, therefore reproducing the effects of Minimal Attachment and Late Closure. However, the results from the naturally occurring text show that the model has learned considerably more than Minimal Attachment and Late Closure. Exactly what the network has learned is an empirical question that we are addressing in current research using statistical analysis (see Costa, Frasconi, Lombardo, Sturt, & Soda, 2002 for some results on these questions).

The results show that there is a viable alternative to modelling experience-based ambiguity resolution in terms of counting explicitly defined configurations (Mitchell et al.,

1995). Apart from the necessity to handle sparse data, there are also methodological reasons why it is preferable to employ a model which does not rely on pre-defined configurations. Throughout the discussion of the psycholinguistic examples, we have seen that the results of frequency searches can differ depending on the amount of detail that is included in the search configurations. In the case of the PP attachment ambiguity, the direction of the preference even reversed when a less detailed configuration was applied. This could lead to the danger of a frequency-based theory becoming unfalsifiable, with the theorist revising the configurations in response to each piece of falsifying data. Thus, it is desirable to employ a mechanism like ours, in which the relevant configurations are not explicitly proposed by the researcher, but are "discovered" by the model itself. Of course, we must acknowledge that, even in the present model, at least some assumptions must inevitably be made about the granularity of representations. For example, by using the Penn Treebank syntactic structures and part-of-speech tags, we are making certain assumptions about the granularity of representations involved.

The results also show that there is a viable alternative to approaches based on probabilistic context-free grammars. This allows for the possibility of modelling preferences such as recency phenomena, which are more global than those that can be characterized in context-free-based systems. It is certainly true, however, that systems based on probabilistic context-free grammars are at a more advanced stage of development than the system proposed here. For example, in the field of computational linguistics, probabilistic context-free grammars combined with methods for calculating lexical dependencies have been extremely successful in parsing real text (Collins, 1996, 1997), and the work of Jurafsky (1996) and Crocker and Brants (2000) shows that lexically-based preferences can be integrated into context-free models in a psychologically interesting way, which is not currently possible in the model described in the present paper. Future work on our model will explore the possibility of incorporating lexical information into the network architecture.

We believe that it is only by scaling up models to deal with wide-coverage language that we can gain a realistic idea of the role of structural frequencies in ambiguity resolution. This is particularly so when we consider competition-based mechanisms, such as those which are proposed in the multiple constraints approach (McRae et al., 1998; Spivey & Tanenhaus, 1998). These models typically lack a mechanism for automatically generating the alternative analyses that are evaluated. However, in competition models, the dynamics of processing are affected by the number of alternatives to the preferred analysis, and their relative activations. Clearly, without a realistic idea of how many alternatives there are, and their relative activations, it is impossible to determine how the competition mechanism should behave. As we have seen in the discussion of the psycholinguistic examples, the number of alternative analyses is often greater than what is standardly supposed by psycholinguists. The model described in the present paper finds these alternatives, and estimates their structure-based probabilities. Thus, it can be used to give a sharper and more realistic test of the predictions of current sentence-processing theories.

The results from this phase of the model development are encouraging for the future implementation of a full parser. In order to build a complete incremental parser with a dynamic grammar and network informer we need to know more than how first-pass decisions are taken. The major topic of research is the implementation of a suitable

recovery mechanism. One possible solution which we intend to test is to use the same network to evaluate competing incremental trees for restarting the analysis, and to examine the data in order to design a more accurate recovery model (Fodor & Ferreira, 1998). Another alternative is to maintain the ranked alternatives, and to re-use them when the current preferred analysis becomes untenable. This would allow a test of the feasibility of models based on ranked parallelism and beam-search (Gibson, 1991, 1998).

Acknowledgements

We are grateful to two anonymous reviewers and to Dan Jurafsky for comments on a previous draft. Various parts of this research have been presented at the AMLaP conference in Leiden, Netherlands, and at the CUNY sentence processing conference in Philadelphia. The research was supported by a British Academy Postdoctoral fellowship.

References

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

Altmann, G. T. M., van Nice, K. Y., Garnham, A., & Henstra, J.-A. (1998). Late closure in context. *Journal of Memory and Language*, *38*, 459–484.

Bader, M., & Lasser, I. (1994). German verb-final clauses and sentence processing. In C. Clifton, L. Frazier & K. Rayner (Eds.), *Perspectives on sentence processing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bever, T. (1970). The cognitive basis for linguistic structures. *Cognition and the development of language* (pp. 279–360). New York: Wiley.

Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. J., & Urbach, T. P. (1995). Syntactic priming – investigating the mental representation of language. *Journal of Psycholinguistic Research*, *24*, 489–506.

Brants, T., & Crocker, M. W. (2000). Probabilistic parsing and psychological plausibility. *Proceedings of the 18th International Conference on Computational Linguistics*. San Francisco, CA: Morgan Kaufmann.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1* (3), 372–381.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. *Proceedings of the 35th annual meeting of the Association for Computational Linguistics* (pp. 16–23). San Francisco, CA: Morgan Kaufmann.

Collins, M. J. (1996). A new statistical parser based on bigram lexical dependencies. *Proceedings of the 34th annual meeting of the Association for Computational Linguistics*. San Francisco, CA: Morgan Kaufmann.

Costa, F., Frasconi, P., Lombardo, V., & Soda, G. (in press). Towards incremental parsing of natural language using recursive neural networks. *Applied Intelligence*.

Costa, F., Frasconi, P., Lombardo, V., Sturt, P., & Soda, G. (2002). Enhancing first-pass attachment prediction. *Proceedings of the 15th European Conference on Artificial Intelligence, Lyons* (pp. 509–512). Amsterdam: IOS Press.

Costa, F., Frasconi, P., Sturt, P., & Lombardo, V. (2002, March). *Exploring the effect of experience on a recursive neural network model of structural preferences*. Paper presented at the 15th annual CUNY Conference on Human Sentence Processing, New York.

Crocker, M. W., & Brants, T. (2000). Wide coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, *29*, 647–669.

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: restrictions on the use of the late closure strategy in Spanish. *Cognition*, *30*, 72–105.

Cuetos, F., Mitchell, D. C., & Corley, M. M. B. (1996). Parsing in different languages. In M. Carreiras, J. E.

García-Albea & N. Sebastián-Galles (Eds.), *Language processing in Spanish* (pp. 145–187). Hillsdale, NJ: Erlbaum.

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

Fodor, J. D. (1998). Parsing to learn. *Journal of Psycholinguistic Research*, *27*, 339–374.

Fodor, J. D. & Ferreira, F. (1998). Reanalysis in sentence processing Dordrecht: Kluwer.

Frasconi, P., Gori, M., & Sperduti, A. (1998). A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, *9*, 768–786.

Frazier, L. (1978). *On comprehending sentences: syntactic parsing strategies*. PhD thesis, University of Connecticut, Storrs, CT.

Frazier, L. (1987). Sentence processing: a tutorial review. In M. Coltheart (Ed.), *Attention and performance* (pp. 559–586). 12. Hillsdale, NJ: Erlbaum.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*, 58–93.

Gibson, E., Pearlmutter, N., Canesco-Gonzalez, E., & Hickok, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition*, *59* (1), 23–59.

Gibson, E., Schuetze, C. T., & Salomon, A. (1996). The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research*, *25*, 59–92.

Gibson, E. A. F. (1991). *A computational theory of human linguistic processing: memory limitations and processing breakdown*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

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

Gilboy, E., Sopena, J. M., Clifton Jr., C., & Frazier, L. (1995). Argument structure and association preferences in Spanish and English compound NPs. *Cognition*, *54*, 131–167.

Goller, C., & Kuechler, A. (1996). Learning task-dependent distributed structure-representations by back-propagation through structure. *Proceedings of the IEEE International Conference on Neural Networks* (pp. 347–352). Washington, DC: IEEE.

Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, *19*, 103–120.

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

Kamide, Y., & Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. *Language and Cognitive Processes*, *14*, 631–632.

Kempson, R., Meyer-Viol, W., & Gabbay, D. (2000). *Dynamic syntax: the flow of language understanding*, Oxford: Blackwell.

Kennison, S. M. (2001). Limitations on the use of verb information during sentence comprehension. *Psychonomic Bulletin and Review*, *8*, 132–137.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, *2* (1), 15–47.

Konieczny, L., Hemforth, B., Scheepers, C., & Strube, G. (1997). The role of lexical heads in parsing: evidence from German. *Language and Cognitive Processes*, *12*, 307–348.

Lombardo, V., & Sturt, P. (2002). Incrementality and lexicalism: a treebank study. In S. Stevenson & P. Merlo (Eds.), *The lexical basis of sentence processing: formal, computational and experimental issues* (pp. 137–155). Philadelphia, PA: John Benjamins.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, *19*, 313–330.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*, 522–533.

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

Milward, D. (1994). Dynamic dependency grammar. *Linguistics and Philosophy*, *17*, 561–605.

Mitchell, D. C. (1989). Verb guidance and other lexical effects in parsing. *Language and Cognitive Processes*, *4* (3), 123–154.

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1995). Exposure-based models of human parsing:

evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, *24*, 469–488.

Phillips, C. (1996). *Order and structure*. PhD thesis, Department of Linguistics and Philosophy, MIT, Cambridge, MA.

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: an eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24* (4), 940–961.

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

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46* (1–2), 77–106.

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

Rayner, K., & Frazier, L. (1987). Parsing temporarily ambiguous complements. *Quarterly Journal of Psychology*, *39A*, 657–673.

Roark, B., & Johnson, M. (1999). Efficient probabilistic top-down and left-corner parsing. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics* (pp. 421–428). San Francisco, CA: Morgan Kaufmann.

Rohde, D. L. T. (2002). *A connectionist model of sentence comprehension and production*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

Schuetze, C., & Gibson, E. A. (1999). Argumenthood and English prepositional phrase attachment. *Journal of Memory and Language*, *40* (3), 409–431.

Sperduti, A., & Starita, A. (1997). Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, *8*, 714–735.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24* (6), 1521–1543.

Spivey-Knowlton, M., & Sedivy, J. C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition*, *55*, 227–267.

Stabler, E. P. (1994). The finite connectivity of linguistic structure. In C. Clifton, L. Frazier & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 303–336). Hillsdale, NJ: Lawrence Erlbaum.

Steedman, M. J. (1989). Grammar, interpretation and processing from the lexicon. In W. M. Wilson (Ed.), *Lexical representation and process* (pp. 463–504). Cambridge, MA: MIT Press.

Stevenson, S. (1994). Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research*, *23* (4), 295–321.

Traxler, M. J., Pickering, M. J., & Clifton, C. (1998). Architectures and mechanisms that process prepositional phrases and relative clauses. *Journal of Memory and Language*, *39*, 558–592.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints on sentence processing: separating effects of lexical preference from garden paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19* (3), 528–553.

Tugwell, D. (1998). *Dynamic syntax*. PhD thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh.

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