Predicting Garden Path Sentences*

ROBERT WILLIAM MILNE

Department of Electrical Engineering
Air Force Institute of Technology
Wright-Patterson Air Force Base
Ohio 45433

This work is an investigation into part of the human sentence parsing mechanism (HSPM). The major test of the psychological validity of any model of the HSPM is that it fail on precisely those sentences that humans find to be garden paths. It is hypothesized that the HSPM consists of at least two processes. We call the first process the syntactic processor, and the second will be known as the semantic processor.

It is hypothesized that the syntactic processor is unconscious, deterministic and fast, but limited. While most ambiguities are resolved on the basis of syntactic information, when the syntactic processor can no longer guarantee a correct analysis, semantic information is used to help resolve the ambiguity. This model leads to a better prediction and explanation of which sentences will cause people to garden path.

GARDEN PATH SENTENCES

A garden path sentence is one which seems to lead people "down the garden path". That is, a person seems to wrongly analyze a portion of the sentence and then, because of later evidence, must go back and reanalyze, or at least correct the mis-analysis.

Marcus (1980, p. 202) says garden path sentences are sentences:

which have perfectly acceptable syntactic structures, yet which many readers initially attempt to analyze as some other sort of construction, i.e., sentences which lead the reader "down the garden path".

The following is a classic garden path:

{1} The horse raced past the barn fell.

*The research reported in this paper is extracted from the author's unpublished PhD thesis titled "Resolving Lexical Ambiguity in a Deterministic Parser". This research was conducted in the Department of Artificial Intelligence and the School of Epistemics, Edinburgh University, Scotland. The author thanks Alan Bundy, Henry Thompson, and two anonymous reviewers for comments and helpful suggestions concerning this paper.
In each sentence of this type, there is a point where two possible analyses are possible, i.e., at "raced". The need to backtrack is a result of selecting an analysis differing from that demanded by the rest of the sentence. For each garden path sentence there is a corresponding sentence that does not require backtracking, e.g.,


This non-garden path partner has the same two possible readings at the same point, but the analysis selected is that demanded by the rest of the sentence. Such a pair of sentences will be called a pair of potential garden path sentences. Each potential garden path sentence may or may not cause a garden path. Of the pair of potential garden path sentences, one is a garden path and the other is not, although which is the garden path is variable. Examples of potential garden path sentences will be marked by curly brackets.

Why do these sentences cause problems for people? Crain and Coker (1979) note: "Bever as well as Chomsky and Lasnik have argued convincingly that unacceptability of GPs is due to processing difficulty." They say "garden path sentences result from the omission of all syntactic markers which signal that one is parsing a Complex NP". This explanation suggests that all garden path sentences should be a problem of un-marked relative clauses.

For example, in {1}, the relative clause marker has been omitted. Other explanations for the difficulty come from Fodor, Bever, and Garret, (1974).

Bever says "the first N. . . V. . . (N) clause. . . is the main clause, unless the verb is marked sub-ordinate" (Bever 1970, Strategy B, p. 294). Fodor, Bever, and Garret (1974, p. 356) have the Canonical Sentoid Strategy to account for the unacceptability of GPs. This is a structure independent mapping from the surface syntactic structure to the semantics. This strategy always "takes the verb which immediately follows the initial NP of a sentence as the main verb, unless there is a surface structure mark of an embedding".

These explanations account for the difficulty in {1}, but, again, suggest that all garden paths are due to the difficulty of an un-marked relative clause.

Wood's definition of a garden path is this:

In human parsing, there are clearly cases where, on the basis of local context and the history of the sentence up to a point, a decision is made to follow a particular alternative and all other alternatives are left to be processed later. This type of processing gives rise to the so-called "garden path" sentence in which the listener is fooled into a false choice among syntactic alternatives and must consciously undo this choice after detecting an inconsistency. (Woods 1973, p. 133)
This definition, like Marcus's, is more general and allows for non-relative clause type garden path sentences. While these definitions seem to account for the examples presented here, do they truly account for all garden path sentences?

It has been suggested (Coker & Crain 1979, Crain & Coker 1979), that the case of the reduced relative garden path, as in {1}, is decided on a semantic basis. They considered examples such as:

{3} The students instructed about the exams were confused.
{4} The professors instructed about the exam were confused.

They felt that one of these was a garden path, and the other was not. They felt that the reader made the verb the main verb of the sentence when the initial NP would fit the subject slot of the verb. A garden path would result whenever this verb should have been the start of a relative clause.

To test their theory, they performed the following experiment. The example sentences were presented in monotone through earphones to the subjects, who were then asked to judge the "truth value" of a possible paraphrase. Reaction times were collected for the judgement on the paraphrase to be made. Examples were of the above form, all of which were intended to be full sentences. If no semantic decision was to be made, then they felt the subjects would garden path on most of the examples. If the decision was made on a semantic basis, then only half of the examples would be garden paths.

Their results showed that only about half of these examples (46%) did cause garden paths. This indicates that the explanations of Fodor, Bever, and Garret above are inadequate.

Their experiment showed that {3} was not a garden path sentence, while {4} was. They then concluded that the subject perceives whether the initial NP will fit the subject slot of the verb. If it will, the subject would accept the verb as the main verb and continue. If however, the initial NP did not fit the subject slot, then the subject would make it a reduced relative clause. The subject would garden path if this semantic decision led to the wrong choice.

Crain and Coker suggest that it is common to discuss horses racing and hence "raced" is accepted as a main verb. Professors instruct normally, but students get instructed, so Crain and Coker suggest that "instruct" will be considered a main verb when used with "professors" more often than with "students". They also predict that people do not garden path on both of {3} and {4}, nor both of the following:

{5} The tenant delivered junk mail threw it in the trash.
{6} The postman delivered junk mail threw it in the trash.

(from Crain & Coker 1979)
According to Crain and Coker's theory, not all potential garden paths will be actual garden paths, but only some of them. The prediction of which will be garden paths is based on the semantic "fit" of the NP into the subject slot.

**MARCUS'S WORK**

**The Definition of Deterministic Parsing**

Central to this paper is the notion of determinism and deterministic parsing. In the following sections, I will explain what deterministic parsing is and the techniques that are used to implement it in a parser.

Marcus (1975) first proposed that English could be parsed deterministically. He stated:

> There is enough information in the structure of natural language in general, and in English in particular, to allow left-to-right deterministic parsing of those sentences which a native speaker can analyze without conscious effort. (Marcus 1980, p. 204.)

I agree with this statement, but the terms "deterministic parsing" and "without conscious effort" must be explained. In relation to the former, Marcus said:

> Natural language can be parsed by a mechanism that operates "strictly deterministic" in that it does not simulate a non-deterministic machine. (Marcus 1980, p. 11.)

Marcus did not propose that deterministic parsing implies that natural language could be parsed by a deterministic machine in the automata theoretic sense. He points out that any computational mechanism that physically exists is deterministic in this sense. The key point of the above statement is that the parser does not simulate a non-deterministic machine. There are two ways a grammar interpreter using a seemingly deterministic grammar can simulate non-determinism. These are backtracking and pseudo-parallelism.

We can prohibit backtracking by insisting that all grammar substruc-tures are permanent. In a parsing context this means that, if one item is attached to another, this attachment can never be broken. i.e., if a PP is attached to an NP, then the parser cannot break the attachment and attach the PP to, say, the VP. If a word is disambiguated to a certain part of speech, it can never be changed to a different part of speech; i.e., if the parser disambiguates "block" as a noun, it cannot change it to a verb. This prevents the grammar interpreter from pursuing a guess that turns out to be incorrect.

It is possible to avoid backtracking, but simulate non-determinism, by taking all possible paths from a given state simultaneously. This is known as pseudo-parallelism.
This method, however, is still not permissible for a deterministic parser. Using pseudo-parallelism, it is possible to follow each permissible transition simultaneously. If one of the paths fails, the parser does not return to a previous state, but, instead, “throws away” any structure built and then terminates that path. This technique is therefore also disallowed. In deterministic parsing, building a constituent and then “throwing it away” is not permitted.

**Conscious Effort**

I do not claim that the entire HSPM is deterministic. I have hypothesized that only the “syntactic processor” is deterministic. Too little is known about the rest of the HSPM to make any statement about it; nor do I claim that all sentences are parsed without conscious effort. I do claim that those sentences which can be parsed by people with no conscious effort, can be parsed deterministically by the “syntactic processor”, even though the semantic processor may assist.

It is assumed that the syntactic and semantic processes of the HSPM rarely fail. When they do, conscious effort is used to recover from the error and the error component takes control. If we assume that every failure of these processes causes conscious effort, we know exactly when a failure occurs. If the model fails at the point where people exert conscious effort, then it will be failing at the same point as the syntactic and semantic processes of the HSPM.

What does “no conscious effort” mean and how do we decide which sentences fall into this category? Unfortunately, the answer to this question is not simple.

It is assumed that the processing of a normal sentence does not require conscious effort and is generally agreed that to understand a garden path sentence requires conscious effort. The reader notices a mental “jump” or “block” when reading of the sentence stops and the garden path is consciously realized. Experimentally, conscious effort is detected by an increase in reaction time to a given task. As an armchair definition; any grammatical sentence that seems abnormal to read, requires conscious effort.

Since the line between conscious effort and no conscious effort is unclear, we will concentrate on those examples that clearly require conscious effort. Without a clear definition and understanding of conscious effort, it is impossible to evaluate deterministic parsing for all sentences in a language. More experimental data must be collected in many areas before we can conclude what does and what does not require conscious effort. Throughout this paper, I will point out when we are assuming that a sentence or fragment requires conscious effort, when we know that it does, and when the deterministic parser predicts that it must.
How is Determinism Accomplished?

One simple technique that makes determinism possible is limited lookahead. "Lookahead" means looking ahead in the input stream before deciding which grammar rule to execute and hence, which will be the next state. This lookahead is always to the next K constituents after the item the parser is currently constructing. (where K may vary from parser to parser). The grammar of this parser is written such that each rule can examine the features of two "buffers" containing constituents.

If it were possible to look arbitrarily far ahead, then it would be possible to have a buffer cell containing each word in the sentence. A single grammar rule which constructed the correct parse tree could then be written for each possible sentence. While a parser that did this would be deterministic, it would not be psychologically plausible. Unrestricted lookahead would also enable the grammar interpreter to simulate a non-deterministic machine by allowing it to perform "closet backtracking".

"Lookahead" means that, if the parser is unsure of a situation, it does not make a random guess. Instead it waits until it has enough information to make the decision correctly. Marslen-Wilson's comment on "securely" indicates that there may be a psychological basis to this approach:

The general claim I want to make is that the human speech understanding system is organized in such a way that it can assign an analysis to the speech input at the theoretically earliest point at which the type of analysis in question can be securely assigned. What is meant by the term "securely" here is that the system does not, within limits, make guesses about the correct analysis of the input (Marslen-Wilson 1980, p. 16.)

My claim is; this applies to the syntactic processor as well as the speech recognition system. Not making guesses is central to the theory of deterministic parsing. By using limited lookahead, it is possible to analyze correctly many sentence forms, without the need for backtracking.

There is one more important strategy in implementing the deterministic parser. A non-deterministic parser is typically driven top-down. It often tries to start a new constituent before it has seen whether any of the following input words can start that constituent. Whenever the parser discovers that there are no lexical items for that constituent, it must backtrack and attempt to find an alternative constituent. Creating a new node before one sees if there are any lexical daughters for it causes much backtracking. Creating a PP node at the end of a NP, before the parser has checked to see if the next word is a preposition, is an example of this guessing. By using lookahead this is not necessary. It is possible to check that the appropriate lexical items are present before a constituent is initiated. My parser does not create a new node until it has a lexical daughter to begin the construction of constituents. One can see that creating new nodes without regard for the
following items is the same as the guessing to which Marslen-Wilson alluded above.

Two Deterministic Parsers

In this section, Marcus’s parser, its motivation and structure will be described. In parallel the Milne parser, ROBIE, will also be described and contrasted with PARSIFAL. The reader who is interested in deterministic parsing is encouraged to read (Marcus, 1980) for a better introduction. The reasons for the modifications of PARSIFAL incorporated in ROBIE will be explained in the subsequent sections.

Marcus first became interested in deterministic parsing when he watched an ATN parser needlessly backtrack, making the same error over and over again, in a situation where the correct solution was obvious to him. He then designed a deterministic parser. To be deterministic, it seemed that the parser would need at least the following three properties:

1. It must be, at least partially, data-driven.
2. It must reflect expectations.
3. It must have some sort of limited lookahead.

These points are fully motivated in Marcus’s work and I will not repeat the motivation here.

The Marcus paper, PARSIFAL, has two main data structures and a grammar interpreter. The first of these is a push-down stack which contains incomplete constituents. This is called the Active Node Stack. The Active Node Stack is constrained such that the constituents it contains must be dominated by a non-terminal node, i.e., a partially built NP, VP, PP, S, etc. If it could contain a terminal (word) that was not dominated by a non-terminal node, then it could be used as an extension of the buffers. This would provide arbitrarily long lookahead, which is forbidden. The Active Node Stack in ROBIE is identical to that used in PARSIFAL.

The other structure is a system of buffers. These are a number of cells containing words or items that have been constructed, but whose grammatical role is unknown. While the buffers can contain any constituent under a single node, for reasons that will be explained below, when a NP is being parsed, the buffers only contain words.

Marcus allowed from three to five buffer cells in his parser and these provided the lookahead capacity. For reasons not relevant to our discussion here, ROBIE has two static buffers. This number is invariant and this fact constitutes one of the major differences between these two parsers.

Both parsers move from left to right over the input string building structure as they proceed. The parsers may suspend construction of a con-
stituent by pushing another item onto the Active Node Stack. The buffers are always the rightmost nodes under consideration by the parsers. They can be considered to be "below" the Active Node Stack, with items being "dropped" from the Active Node Stack into the buffers.

In ROBIE, these buffers are fixed always to be the next two cells after the bottom of the Active Node Stack. The buffers move right and left as items are pushed onto and popped off of the stack. The grammar and the rule matcher can only look at two buffers. (The one exception is conjunction. No one has yet researched a satisfactory conjunction method for a deterministic parser. I know of no solution that can handle a wide range of conjunctions without a special mechanism.)

In parsing the sentence "The shy boy has kissed Mary.", the state of ROBIE, which is very similar to the state of PARSIFAL, would be as follows:

Packet: CPOOL
Rule about to run: PROPNAME pattern: [name]
Active Node Stack:
top: S-NP the [SS-FINAL, CPOOL]
shy
boy
AUX has
bot: VP kissed [SS-VP, CPOOL]
Buffers: [Mary][.]

As words are first considered by ROBIE, they arrive into B2, the second buffer. For PARSIFAL words arrive into whichever buffer is the rightmost. In both parsers, the grammar may indicate the start of a new constituent, based on the syntactic features of the next word. This constituent will be placed on the bottom on the Active Node Stack. The Active Node Stack grows downward and only the bottom item of the stack is currently Active i.e., the only item being actively constructed at the time. In this example, the VP node is the Current Active Node and work on the S node has been suspended until the VP node is finished.

In the diagram, the first buffer contains the word "Mary" and the second buffer contains the word "."

In both parsers, the grammar consists of a set of production system rules ordered into packets. Each rule consists of a pattern for the head, that serves for the production system pattern, and a body that is executed by the interpreter once the rule has been selected. Every rule has a name, a priority and is a member of a packet.

In both parsers, only rules in active packets can be tested by the interpreter. The body is not touched until the rule has matched. A rule body can activate and deactivate packets. This provides the top-down component of PARSIFAL and ROBIE.
In PARSIFAL, the patterns of the grammar rules could match any combination of the contents of the three buffers and the bottom of the Active Node Stack. (The item currently being built). PARSIFAL could also check features on the lowest S node of the stack. This means that each pattern could inspect up to five nodes before matching. In ROBIE, the patterns are constrained to match only two buffers and it is not possible to access any node in the Active Node Stack except the bottom one. The reasons for this will be described in the next section.

The grammar rules in ROBIE are very similar to Marcus's original rules. In fact, most of the rules have the same names. They have, however, been modified slightly, as will become apparent later.

The production system grammar rules were structured by combining groups of rules into packets; a packet being a collection of rules. These packets can be made active or inactive by the parser. Only active rules can match against the state of the parser. Each node on the Active Node Stack has a list of active packets associated with it. In the above diagram, the packets SS-VP and CPOOL are associated with the VP node, and the packets CPOOL and SS-FINAL are associated with the S node. If the node is not the Current Active Node (the bottom most node), then the packets associated with it are inactive.

We can now look at the three desirable properties mentioned above and see how both parsers meet them.

1. The parsers are data-driven in that the patterns of the rules will match on the words as they arrive in the buffers, or other items in the buffers.
2. The parsers reflect expectations since only rules in active packets can match. By activating the packets for the expectations we have, the parser is providing a top-down component.
3. The buffers provide a constrained look ahead. Before a rule can match, the parser can examine the contents of a limited number of buffers.

**NOUN/VERB AMBIGUITY**

Let us look at a simple case of ambiguity and how it can be resolved with a two buffer lookahead. We will then turn to a more difficult example. The sentence fragment:

[7] The toy rocks....

could be completed as either:

{8} The toy rocks are red.
{9} The toy rocks easily.
In sentence {8} the subject NP is "the toy rocks" and "rocks" is a noun, while in {9} the subject NP is only "the toy" and "rocks" is a verb. In order to find the end of the noun phrase properly, the parser must detect these possibilities and decide which is applicable. A non-deterministic parser with a backtracking capability, could always try one analysis first. If this fails, it can try the other analysis by backtracking. The only difficulty is which alternative to try first as a matter of efficiency.

A deterministic parser however, is not able to backtrack and hence cannot follow this strategy. Once the deterministic parser "decides" that a word is a noun, it is committed and cannot change its mind. Hence, the deterministic parser must be able to decide what part of speech the ambiguous word is without making an error.

My deterministic parser uses two buffer lookahead to see the following word and determine the part of speech of the word. For example, when parsing {8}, the parser would have the word "rocks" in the first buffer and the word "are" in the second buffer, i.e.:

[rocks] [are]

This pattern indicates that "rocks" was being used as a noun in this sentence. For {9}, the buffers would be:

[rocks] [easily]

showing the verb usage. By looking at the following word, the parser can handle these examples without needing to backtrack.

Using two buffers then, it is often easy to decide which part of speech a word is being used as without needing to backtrack. This is a much better result than the non-deterministic approach explained above.

The Active Node Stack cannot contain words unless they are dominated by a non-terminal node, i.e., a partially built constituent. Because of this, the lookahead buffers will always contain words when the headnoun of a NP is being constructed. The detailed reasons for why this must be true are not relevant to our discussion here, but will become clear later in the text.

Let us now look at a more difficult example of this ambiguity. In this case, the two buffer lookahead is not sufficient to resolve the ambiguity. In these examples, the second buffer is not sufficient to disambiguate the ambiguous word properly. To distinguish between {10} and {11}, it seems we would need four buffers and to distinguish between {12} and {13} we would need eight buffers since we need to see the word "faded" before we could resolve the ambiguity. If the word "block" is currently in the first
buffer, we need six buffers containing "the sun shining on the house", plus a buffer containing "faded". This is eight buffers in total.

In fact, there can be an arbitrary number of words between "blocks" and the word that indicates whether "blocks" is a noun or a verb (see Milne, 1978). Therefore, no fixed amount of lookahead will be able to disambiguate the word "blocks" in this situation. In no way could a two buffer deterministic parser handle these examples. As was explained in the previous section, the use of an arbitrary number of buffers would make our claim of psychological plausibility vacuous.

The above sentences are in two pairs of potential garden paths. Within each pair, the sentences are the same for most of the string, but differ at the point at which the function of the word in question (blocks) can be ascertained (at faded). For the parser this means that the buffers will contain the same items and the disambiguating word will be beyond the buffers (to the right). \{10, 11\} and \{12, 13\} form two such pairs. For all of these, the buffers contain \[blocks\] \[the\] and the disambiguating word is too far to the right.

Milne, (1978) explored handling noun/verb ambiguity in a deterministic parser and showed that the "building blocks" sentences could lead to garden paths and hence each one is a potential garden path. In fact this paper showed that all situations involving a word that could be either a noun or a verb followed by a word that could be a noun or a verb and was plural, could lead to a garden path. The above definitions do not explain why these could be garden path sentences. Milne, (1978) made no attempt to handle these situations.

Consider the following sentences:

\{14\} The toy rocks near the child quietly.
\{15\} The toy rocks near the child are pink.

These are a pair of potential garden path sentences. They help to demonstrate that, in the situation of a singular noun/verb word followed by a plural noun/verb word, it is always possible to finish the sentence so that it will be a garden path. We will refer to this type of a garden path as a plural garden path.

As previously stated, the lookahead may need to be arbitrarily long to handle these examples. In these cases, the deterministic parser cannot disambiguate all of them properly.

At this point, we must return to the motivation for the parser. It was constructed to parse in a psychologically plausible way. This means that we want the parser to perform in exactly the same way as people do. If the parser were to parse sentences that people found incomprehensible, then it would not be fulfilling its role as a psychological model of normal human sentence processing.
We have noted that, by definition, people fail on garden path sentences, needing to employ conscious effort in their analysis. The parser, then, should also fail whenever it is presented with a garden path. This point was made by Marcus, who stated that any deterministic parser should fail on this type of sentence. He says:

> a deterministic parser ideally should take the garden path and become “stuck” at exactly the point at which people become conscious that they have been misled. (Marcus, 1980, p. 204.)

While his parser was not built to handle garden path sentences, it did fail on this type of sentence. Not all garden path sentences, however, caused his parser to fail.

In order to limit the model to successfully parsing only non-garden path sentences, those sentences on which we wish the parser to fail must be identified. Since garden paths are defined as sentences on which people fail, this can be done by experimentation.

Assuming that the parser is limited in such a way, it can itself be used to predict which sentences will cause problems in humans and, hence, give a definition of a garden path sentence as one which cannot be deterministically parsed by the model. The following experiments test the model against human performance, in its ability to distinguish and fail on garden path sentences.

### The Garden Path Prediction of PARSIFAL

Let us look at the garden path prediction of Marcus's parser. His parser consisted of a Active Node Stack, where partially built items resided, and three buffers. Each buffer contained a word or constituent that could be represented by a single node. A buffer can then hold a word, NP, PP, VP, etc. (We will use a version of X bar theory (Postal, 1974). NP stands for Noun Phrase, VP for Verb Phrase, PP for Prepositional Phrase and S for a sentence. S̅ is an embedded S, toVP is an embedded VP with the auxverb “to”.) In the best situation, an ambiguous word will be in the first buffer and the lookahead will be two items (Buffers 2 and 3). When an NP is being built, these two items of lookahead will be words and never whole NPs or larger items.

At the time the word “rocks” is being analyzed in the fragment “the toy rocks...”, the parser's state would be as below: (The symbol [noun], will mean a buffer that contains an item with the syntactic feature “noun”, symbols such as [rocks] will mean a buffer that contains the word rock.)

Active Node Stack: NP the toy
Buffers: [rocks] [the] [child]
A sentence is predicted to be a potential garden path if the lookahead is insufficient to disambiguate the word correctly. Marcus did not recognize the concept of a potential garden path sentence. Instead of the pair of sentences we have called potential garden path sentences, he considered one a garden path and the other an ordinary sentence.

For the examples in this section, the lookahead is not sufficient to resolve the ambiguity. We have said that each time a person encounters a pair of potential garden path sentences, one and only one is a garden path. Marcus’s prediction says that one will be a garden path, but cannot tell which it is.

Because of this, his parser would arbitrarily choose one case (i.e., the noun usage) to be the preferred usage. For example in the case of a word that could be a plural noun or a singular verb, his parser would always choose the plural noun usage. This would be correct for:

{16} The granite rocks by the seashore are eroded.
{17} The granite rocks by the seashore with the waves.

In {16}, “rocks” is used as a noun, and Marcus’s approach would correctly predict {17} to be a garden path. But in:

{18} The statue stands in the park.
{19} The statue stands in the park are rusty.

where “stands” is used as a verb in {18}, his approach would incorrectly predict {18} to be a garden path and not {19}. Therefore, this approach would predict some sentences to be garden paths that are not.

Marcus’s prediction of garden path sentences also tells us how “short” a garden path must be. In sentence {12}, the disambiguating word is outside the three buffers. What happens when the word is in one of the three buffers? Marcus’s prediction would say that it was not a garden path. If the parser used all the information in the three buffers, then for a sentence to be a garden path, there would need to be at least three words between the ambiguous point and the disambiguating word. Are all garden path sentences this long?

Why is it Wrong?

The well-known garden path:

{20} The prime number few.

is a counter example to Marcus’s prediction. When PARSIFAL is analyzing the word “number” the state will be:

Active Node Stack: NP the prime
Buffers: [number] [few] [.]
Since the entire sentence fits into the three buffers, all information to analyze the sentence is available. But people do garden path on this sentence. I claim the following are also counter examples:

{21} The granite rocks during the earthquake.
{22} The sentry stands are green.

Even though the number of words read, before the error is realized, is very small, people are aware of some confusion while analyzing these sentences. Again, all the information for proper analysis is contained in the three buffers and it is not predicted to be a garden path. I will later provide experimental evidence that these sentences really are garden paths.

This shows that Marcus’s garden path prediction is inadequate. While the prediction will correctly tell us some sentences are garden paths, it will also predict some sentences are garden paths that are not, and judge as acceptable some sentences which are garden paths.

So how can we predict whether a sentence will be a garden path? What is happening that causes the garden path? Can our model be extended to answer these questions?

The New Theory

I feel that the previous definitions are inadequate because they fail to incorporate semantic information. To account for the difficulty of garden path sentences, I propose the following hypothesis to describe what people do when they encounter a potential garden path.

Semantic Checking Hypothesis:

When a person encounters a situation which syntactic context implies might lead to a garden path, instead of using lookahead, they decide which alternative to pursue based on semantic information. They do this without regard to the following words in the sentence. If their preference for this leads to an analysis different from that demanded by remainder of the sentence, they will garden path.

In our example of a noun/verb word followed by a noun/verb word that is plural, a person would attempt to make a complex item name of the two words. The prediction now depends on a semantic preference for noun/noun combinations, rather than on lookahead, for garden path sentences. I will expand this theory more in the following section.

Does This New Explanation Fit Our Examples?

As the above examples show; people like prime numbers, but don’t like sentry stands. I believe people will garden path if “prime number” is not a complex headnoun because it is a common construction, as with “aluminium
screws" and "granite rocks". People will use "stands" as a verb in "sentry stands" since it is more common to have a sentry standing. Finally for the case of toy rocks, both constructions are equally possible, so some people would garden path on the first and some on the second. It is also very easy to bias this last case with context, etc. altering the prediction.

The Semantic Checking hypothesis predicts that

{23} The sentry stands on guard.

will not be a garden path, but

{24} The sentry stands are green.

will be.

This new theory makes definite predictions of what will be a garden path sentence, based on a person's semantic preference for complex head-nouns. Lookahead is predicted to have no affect on these examples, but instead a semantic decision is made. In the next sections, I will present an experiment to test this.

THE FIRST EXPERIMENT

Purpose

The purpose of this experiment was to test the garden path effect. The task used was a simple collection of reading times, as used by Cirilo and Foss 1980; Graesser, Hoffman, and Clark 1980; and Just and Clark, 1973. The subject sat in front of a Visual Display Unit and instructions were presented on the screen of the Visual Display Unit. The subject was then presented with a series of sentences. When he had read and understood each sentence, the subject pressed a key on the keyboard.

A Terak micro computer was used to collect reaction times. The time was measured from the presentation of the sentence until the key was pressed. The next sentence was presented immediately. Twenty-two undergraduate students from the University of Edinburgh participated in this experiment. All were unpaid volunteers. None of the subjects were familiar with the notion of global ambiguity or garden path sentences.

Examples

The examples were arranged in two orders to control for sequence effects. Each subject was given one order and each order contained all sentences, the sequence of the sentences being random. Between the test orders, the sequence of each sentence with its control was reversed. The predictions varied with each group of sentences. These will be discussed with the results.
The subject was given 10 practice sentences before the test sentences appeared. A total of 50 examples were tested in the experiment. The sentences tested that were relevant to this paper are as follows:

**Garden Paths**

[26] The boat floated down the river quietly.
[27] The horse raced past the barn fell.
[28] The horse raced past the old barn.
[29] The cotton clothing is made of grows in Alabama.
[30] The cotton clothing is made in sunny Alabama.

**Results**

A Student's *t*-test was performed on each pair of sentences, on each order and for the combined results. Listed below is the mean time (in seconds) for all sentences. The result are presented in the tables below with the number referring to the previous examples. Because there are so many examples, I will discuss them in their logical groupings. We chose the .01 level of significance for this experiment.

Farther to the right is the result of ROBIE's effort to analyze the same sentence. The phrase "success" means that the parser successfully analyzed the sentence. The phrase "failed at X", means the parser "blocked" or garden pathed at that point.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Mean (seconds)</th>
<th>Rules ran by ROBIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>[25]</td>
<td>9.38</td>
<td>failed at &quot;sank&quot;</td>
</tr>
<tr>
<td>[26]</td>
<td>3.71</td>
<td>success</td>
</tr>
<tr>
<td>[27]</td>
<td>10.13</td>
<td>failed at &quot;fell&quot;</td>
</tr>
<tr>
<td>[28]</td>
<td>3.05</td>
<td>success</td>
</tr>
<tr>
<td>[29]</td>
<td>9.23</td>
<td>failed at &quot;grows&quot;</td>
</tr>
<tr>
<td>[30]</td>
<td>4.04</td>
<td>success</td>
</tr>
</tbody>
</table>

It was predicted that the first sentence of each pair would cause a garden path, but not the second. The first of each pair was significantly different from the second at the .01 level. These sentences proved to be extremely difficult for the subjects. After looking at the garden path examples for several seconds, many subjects asked what to do if they could not understand the sentence! The experiment did not test whether the subject had really analyzed the sentence properly. A few subjects admitted afterwards that they couldn't understand the sentence, so went on to the next sentence. These show conclusively that the first of each pair causes a garden path, while the second sentence of each pair does not.

This experiment has shown that garden path sentences do require a much longer reading time relative to their non-garden path partner. For
most examples the garden path sentence required a much greater reading time.

THE SECOND EXPERIMENT—
TESTING GARDEN PATH SENTENCES

The Semantic Checking Hypothesis makes definite testable predictions and so an experiment was conducted to test them. The purpose of the experiment was to show that, of a pair of potential garden path sentences, one was a garden path and to demonstrate that subjects did not use lookahead to resolve this ambiguity. As our theory in the previous section predicts, for these sentences:

[31] The chestnut blocks are red.
[32] The chestnut blocks the sink.

one will cause a garden path and a longer reaction time, when compared to the other. In both the above sentences, looking just one word ahead is sufficient to resolve the ambiguity as the second buffer contains the disambiguating word. So, a person looking at the second buffer, could resolve the ambiguity and the reaction times should be the same for both sentences.

If the theory is wrong, then responses to both sentences will take the same length of time. If the theory is correct, the response to one of the sentences will take longer than to the other. As the hypothesis states, context may have a strong effect on the understanding of these sentences. Therefore, it is predicted that the sentence of the pair which requires the longer time may vary from time to time for any one person and also from one person to another.

The Pre-Test

First, it was important to decide whether there was a generally preferred reading for the noun/noun combinations which were to be used in the experiment. To collect examples, a written survey was conducted which consisted of 21 fragments, as below. The subject was asked to complete the series of words such that they formed a complete sentence. There were two orders to control for order effects and 50 subjects participated.

The part of speech use of the noun/verb/plural word was then checked. A tally was made of each time the word in question was used as a noun and as a verb. This provided an indication of the semantic preference for these word pairs. For example, if almost all subjects completed “grappling hooks” using “hooks” as a noun, then this combination was considered semantically preferred. If most of the subjects completed “boy screws” us-
ing "screws" as a verb, it was concluded that the verb reading was semantically preferred.

The examples were then divided into three groups. The first group contained pairs that were strongly preferred as noun/noun combinations, the second, pairs which were not preferred as noun/noun combinations and the third group were examples showing an equal split among subjects, or no bias. The results by group are as follows:

<table>
<thead>
<tr>
<th>Noun/Noun Preference</th>
<th>Noun Uses</th>
<th>Verb Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>[31] the grappling hooks</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>[32] the aluminium screws</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>[33] the granite rocks</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>[34] the map pins</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>[35] the top hooks</td>
<td>34</td>
<td>9</td>
</tr>
<tr>
<td>[36] the truck handles</td>
<td>32</td>
<td>12</td>
</tr>
</tbody>
</table>

**Verb Preference**

| [37] the sentry stands | 0 | 44 |
| [38] the boy screws | 0 | 50 |
| [39] the cook handles | 0 | 50 |
| [40] the jeep rocks | 0 | 44 |
| [41] the statue stands | 8 | 42 |
| [42] the sniper pins | 7 | 32 |
| [43] the arm hooks | 8 | 42 |
| [44] the chestnut blocks | 13 | 37 |

**Equal Bias**

| [45] the toy rocks | 21 | 26 |
| [46] the bike handles | 25 | 23 |
| [47] the plastic blocks | 25 | 25 |
| [48] the building blocks | 30 | 20 |
| [49] the cover screws | 20 | 30 |
| [50] the flower stands | 18 | 23 |
| [51] the book pins | 20 | 28 |

**Testing the Predictions**

The experiment was based on the above data and collected reaction times to test the hypothesis that, of the pair of potential garden paths, one is a garden path. The task was to read the series of words presented and decide whether they were a complete sentence, or just a fragment. The noun/noun combinations used were taken from the above list of pairs. The examples were in two groups, those with a strong semantic bias and those with no semantic bias. The first group will be called the "biased examples". Examples were picked that had a strong preference for or against the combination, but could be used as both a noun/noun combination and a noun/verb combination. A sentence was constructed for each combination, using the
combination in the non-preferred way. A partner sentence was then constructed which was matched in syllables and words, which used the combination of words in the preferred usage. For example:

[33] The sentry stands are green.
[34] The sentry stands on guard.

It is predicted that [33] will be a garden path. The pre-test has shown that "sentry stands" is not preferred as a noun/noun combination, although it is used as such in the sentence. According to the theory, the subject will attempt to use "stands" as a verb rather than as part of the headnoun combination. This will lead to the wrong analysis, and a garden path will result. In each sentence, the features of the word following the target word (the word with the features noun, verb, plural i.e., "stands") were sufficient to disambiguate the target word.

It was predicted that [33] would take longer than [34]. However, the difference in reaction time might simply be the result of the different syntactic structures of the two sentences. Therefore, for each sentence of the above pair, a control sentence was constructed. This sentence was matched in syllables, but the word preceding the target word was changed to make the sentence a definite noun use (for example by switching to an adjective); similarly, for the second sentence. This produced controls of the form:

[35] The pencil stands are green.
[36] The army stands on guard.

It is predicted that the control for the non-garden path sentence, [36], would take the same time as the non-garden path member of the potential garden path pair, [34]. To ensure that the longer reaction time for the verb usage was not due to structural and processing differences, the two controls ([35] and [36]), were compared. If they required the same time, then it could be concluded that the verb usage did not require a longer time to process. If the verb usage took longer to process, then it was necessary to check that the effect was greater than would be due to structural differences. If three of the sentences required the same time to read and one of the test sentences required a longer time, then the predicted result would have been achieved.

The above prediction was only valid if the subject judged both examples of the test pair to be a sentence. If the subject judged [34] to be a fragment then, we can conclude the noun reading of "stands" was used in both sentences. In this situation, the reaction time for both sentences ([33], [34]) would be the same.

For each test combination, we now have four sentences. To keep the task valid, four fragments were also included so that the subject was presented with an equal number of sentences and fragments.

The test pairs used were as follows:

[37] The prime number few.
[38] The bold number few.
Each pair above is a pair of potential garden path sentences. The first of each pair was predicted to be a garden path sentence on the basis of the results of the pre-test.

The second test group, known as the "non-biased examples", will not be reported in this paper.

The sentences were presented in two different orders to control for order dependence. The biased examples were presented before the non-biased examples in both orders. Each subject was given an order and all types of sentences were randomly ordered. All subjects were tested on all sentences.

Subjects

Forty-seven undergraduate students from Edinburgh University participated in the experiment. Most were from the Psychology department. All were unpaid volunteers and native speakers of "British English". Approximately half the students were tested on each order.

Procedure

A Commodore Pet micro computer was used to collect reaction times for the subjects. Each person sat in front of the display screen and the instructions were read aloud by the experimenter. The sentences were then displayed on the center of the Visual Display Unit. The subject was asked to decide whether the series of words presented was a complete sentence, or just a fragment. If the subject thought the series of words was a fragment, he pressed a key with his right hand. If he felt it was a sentence, he pressed a key with his left hand. Reaction times were measured from the presentation of the sentence until the response. The subject was told that all series of words would be syntactically and semantically well formed. He was first given 16 practice sentences. Sentences were presented in groups of twenty with a short rest between each group and there were 86 examples in total.
After the test, each subject was asked how he had done. Those reporting they did badly, or had trouble, were noted for later analysis.

**Analysis of the Data**

The reaction times and judgements of all the subjects were checked for irregularities. Remember that for each test pair in the biased examples, 4 sentences and 4 fragments were constructed. The sentence/fragment judgements of each subject were checked for the biased examples and the number of judgements differing from this design were tallied. The average number of different judgements per subject was 5 sentences (10%). Sentences with different judgements were not removed from the analysis. 18% of the subjects had differences in judgement on more than 11 of the sentences. For example, some of these subjects listed all examples as complete and others as all fragments. As the error rate for these subjects was over 20%, they were removed from the analysis.

If the test word was used as the same part of speech in both examples, then no "garden pathing" should have occurred and both examples should require the same time to read. Therefore, if one of the examples was judged to be a fragment, and prediction was that both would require the same length of time. Very few subjects made this altered judgement and their results were not separated from the others. As a result some of the variances are slightly larger than they would be if these people were checked separately. Exceptionally long or short reaction times were regressed to their group’s mean but without losing their significance or relative ordering.

For the "granite rocks" examples, approximately half the subjects considered one of the test sentences a fragment. It was only predicted that one of these would be a garden path when both examples were judged to be complete sentences. In order for this to happen, the noun/verb/plural word would have to be used as a noun in one sentence and a verb in the other sentence. If the example was judged to be a fragment, then it is assumed that the noun/verb/plural word was used as the same part of speech in both sentences.

For the "granite rocks" examples, the responses were split into two groups and the same analysis as above was performed on each group. The time reported below is for the test pair that was predicted to be significantly different. There was no significant difference among the sentences predicted to be the same, \( F < 3 \).

A separate analysis of variance with repeated measures was performed for each set of four sentences in the biased group. (i.e., the G.P., non-G.P. and 2 controls). For the groups that the \( F \) test showed were significantly different, the Newman-Keuls test was used to determine which pairs were significantly different. These results are reported below.
A Student's t-test was performed on each of the test pairs and across their corresponding controls. These results also supported the findings presented below. A further analysis of variance was performed using the model \( \text{Time} = \text{Reading Rate(Person)} \ast \text{Difficulty(Sentence)} \ast \text{Error} \). This model was transformed to \( \log(\text{Rate}) + \log(\text{Difficulty}) + \log(\text{Error}) \). The times for the sentences were fitted against this linear model and the standard error was then checked for significance. A plot of the error was performed which confirmed that it was normally distributed. This model supported the findings presented below. These tests were made on each order and across both orders.

**Results**

The results of the test sentences are shown below. I will present each group of four examples and the mean reaction time in seconds for all examples combined from both orders. The star (*) indicates that the sentence was predicted to take longer. All results are at the .01 level of significance.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Mean Reaction Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[49] The bold number few.</td>
<td>2.22</td>
</tr>
<tr>
<td>[50] The prime number few.</td>
<td>3.15</td>
</tr>
<tr>
<td>[51] The bold number two.</td>
<td>2.54</td>
</tr>
<tr>
<td>[52] The prime integer two.</td>
<td>2.35</td>
</tr>
</tbody>
</table>

All subjects judged [50] to be a fragment. Hence the prediction was that all four sentences would require the same time to read. These examples were not significantly different, \( F(3,102) = 3.7 \).

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Mean Reaction Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[53] The shotgun pins were rusty from the rain.</td>
<td>3.41</td>
</tr>
<tr>
<td>[54] The sniper pins were rusty from the rain.</td>
<td>*3.63</td>
</tr>
<tr>
<td>[55] The sniper pins the victim in the woods.</td>
<td>2.68</td>
</tr>
<tr>
<td>[56] The sniper guards the victim in the woods.</td>
<td>2.54</td>
</tr>
</tbody>
</table>

The reaction times for these sentences were significantly different, \( F(3,102) = 22.4 \). The Newman-Keuls test showed that [54] was significantly different from [55] and [56]. It also showed that [53] was significantly different from [55] and [56]. This supports the prediction that [54] is a garden path, but does not rule out the theory that the time difference is due to a structural processing difference.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Mean Reaction Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[57] The pencil stands are green.</td>
<td>1.88</td>
</tr>
<tr>
<td>[58] The sentry stands are green.</td>
<td>*2.58</td>
</tr>
</tbody>
</table>
[59] The sentry stands on guard. 1.93
[60] The army stands on guard. 1.99

These sentences were significantly different, $F(3,102)=7.3$. The Newman-Keuls test showed that [58] was significantly different from the other three, as predicted.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Mean Reaction Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[61] The wooden blocks are red.</td>
<td>1.79</td>
</tr>
<tr>
<td>[62] The chestnut blocks are red.</td>
<td><strong>2.56</strong></td>
</tr>
<tr>
<td>[63] The chestnut blocks the sink.</td>
<td>1.91</td>
</tr>
<tr>
<td>[64] The stopper blocks the sink.</td>
<td>2.42</td>
</tr>
</tbody>
</table>

These examples were significantly different, $F(3,102)=19.2$. The Newman-Keuls test showed that [62] was significantly different from [61] and [63]. It also showed that [64] was significantly different from [61] and [63]. Although sentence [64] takes longer to process than [61], presumably due to structural differences, the garden path example corresponds to the faster sentence. This rejects the theory that the difference in reaction time is due only to a structural processing differences for these examples.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Mean Reaction Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[65] The table rocks during the earthquake.</td>
<td>3.10</td>
</tr>
<tr>
<td>[66] The granite rocks during the earthquake.</td>
<td><strong>4.18</strong></td>
</tr>
<tr>
<td>[67] The granite rocks were by the seashore.</td>
<td>3.04</td>
</tr>
<tr>
<td>[68] The biggest rocks were by the seashore.</td>
<td>3.71</td>
</tr>
</tbody>
</table>

These examples were significantly different, $F(3,42)=5.6$. The Newman-Keuls test showed that [66] was significantly different from [65] and [67]. Comparison shows that [65] was faster than [68], suggesting that some time difference was due to a structural processing difference. But this difference suggests that [66] should be the slowest sentence, rather than the fastest. These results again support the theory.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Mean Reaction Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[69] The sign pins onto the wall.</td>
<td>3.46</td>
</tr>
<tr>
<td>[70] The map pins onto the wall.</td>
<td><strong>2.74</strong></td>
</tr>
<tr>
<td>[71] The map pins are bright red.</td>
<td>2.44</td>
</tr>
<tr>
<td>[72] The large pins are a bright red.</td>
<td>2.10</td>
</tr>
</tbody>
</table>

These examples were significantly different, $F(3,102)=8.9$. The Newman-Keuls test showed that [69] was significantly different from [71] and [72]. Sentences [70] and [71] were not significantly different at the .01 level. This result was unexpected. The result may be due to a garden path effect in [69] or to structural differences. The experiment does not provide enough data to test these hypotheses.
Discussion

The results support the Semantic Checking Hypothesis. They show that the reaction times were indeed significantly different for the test pairs. This suggests that one of the examples did cause a garden path. However, if Marcus's theory, as applied to humans, was correct then the reaction times should have been more or less the same. Therefore, Marcus's theory is wrong and something else must be happening. This supports the theory that semantic information is used to resolve the ambiguity in these potential garden paths. If the subject had used lookahead, then the reaction times for the test sentences would have been the same. As they were different, it is concluded that lookahead was not used.

It is very hard to test this theory. Since the pairs can be biased by context, then it is easy for a subject to change in the experiment and show no overall effect. Whenever a subject did not understand a test example, he could press "fragment". Many of the subjects considered "the statue stands he saw" as unacceptable and admitted judging it to be a fragment for this reason. Sentences that are semantically odd can also take a longer time to read as the subject attempts to understand the sentence. This experiment does not rule out the possibility that the subject did not garden path on each of the test sentences, and the longer reading time was caused by the subject puzzling over the "semantically odd" portion of the sentence. Finally, the subject may be inconsistent in his reaction to odd sentences, again causing great variance.

CONCLUSION

This work was an investigation into part of the human sentence parsing mechanism (HSPM). We hypothesized that the HSPM consists of at least two processes. We called the first process the syntactic processor, and the second the semantic processor. The syntactic processor is unconscious, deterministic and fast, but limited. The resolution of lexical ambiguity was used as a vehicle to investigate this hypothesis. During processing of some sentences, the syntactic processor, at key points, asked the semantic processor to make a decision in order to resolve an ambiguity. These key points are whenever a general situation arises where the syntactic processor can no longer guarantee a correct analysis, for example in a potential garden path situation. A major focus of this research was the identification of those situations in which people use the semantic processor to assist with the resolution of ambiguity and the sentences in which this occurs. It was shown that these situations were correctly predicted by the two buffer lookahead of the syntactic processor.
Perhaps the most important result of this paper is when the syntactic and semantic processes interact. These two processes interact whenever an ambiguity arises which the syntactic process cannot guarantee to resolve correctly, because of its limitations. This is contrasted with a theory that suggests that the semantic processor is only used when the syntactic processor has failed. The main difference is that they interact when the syntactic processor might fail, rather than when it has failed.

REFERENCES


Milne, R. Using Determinism to Predict Garden Paths. *AISB-80 Conference Proceedings,* Amsterdam, July 1980. (a)

Milne, R. Parsing Against Lexical Ambiguity. *Proceedings of the COLING-80 Conference,* Tokyo, August 1980. (b)


