Conceptual Analysis of Garden-Path Sentences

Michael J. Pazzani
The MITRE Corporation
Bedford, MA 01730

ABSTRACT

By integrating syntactic and semantic processing, our parser (LAZY) is able to deterministically parse sentences which syntactically appear to be garden path sentences although native speakers do not need conscious reanalysis to understand them. LAZY comprises an extension to conceptual analysis which yields an explicit representation of syntactic information and a flexible interaction between semantic and syntactic knowledge.

I. INTRODUCTION

The phenomenon we wish to model is the understanding of garden path sentences (GPs) by native speakers of English.

Parsers designed by Marcus [81] and Shieber [83] duplicate a reader's first reaction to a GP such as (1) by rejecting it as ungrammatical, even though the sentence is, in some sense, grammatical.

(1) The horse raced past the barn fell.

Thinking first that "raced" is the main verb, most readers become confused when they see the word, "fell". Our parser, responding like the average reader, initially makes this mistake, but later determines that "fell" is intended to be the main verb, and "raced" is a passive participle modifying "horse".

We are particularly interested in a class of sentences which Shieber's and Marcus' parsers will consider to be GPs and reject as ungrammatical although many people do not. For example, most people can easily understand (2) and (3) without conscious reanalysis.

(2) Three percent of the courses filled with freshmen were cancelled.

(3) The chicken cooked with broccoli is delicious.

The syntactic structure of (2) is similar to that of sentence (1). However, most readers do not initially mistake "filled" to be the main verb. LAZY goes a step further than previous parsers by modeling the average readers ability to deterministically recognize sentences (2) and (3).

If "filled" were the main verb, then its subject would be the noun phrase "three percent of the courses" and the selectional restrictions [KATZ 63] associated with "to fill" would be violated. LAZY prefers not to violate selectional restrictions. Therefore, when processing (2), LAZY will delay deciding the relationship among "filled" and "three percent of the courses" until the word "were" is seen and it is clear that "filled" is a passive participle. We call sentences like (2) semantically disambiguatable garden path sentences (SDGPs). Crain and Croker [79] have reported experimental evidence which demonstrates that not all potential garden path sentences are actual garden paths.

LAZY uses a language recognition scheme capable of waiting long enough to select the correct parse of both (1) and (2) without guessing and backing up [MARCUS 76]. However, when conceptual links are strong enough, LAZY is careless and will assume one syntactic (and therefore semantic) representation before waiting long enough to consider alternatives. We claim that we can model the performance of native English speakers understanding SDGPs and misunderstanding GPs by using this type of strategy. For example, when processing (1), LAZY assumes that "the horse" is the subject of the main verb "raced" as soon as the word "raced" is seen because the selectional restrictions associated with "raced" are satisfied.

One implication of LAZY's parsing strategy, is that people could understand some true GPs if they were more careful and waited longer to select among alternative parses. Experimental evidence [Matthews 79] suggests that people can recognize garden path sentences as grammatical if properly prepared. Mathhews found that subjects recognized sentences such as (2) as being grammatical, and after doing so, when later presented with a sentence like (1) will also judge it to be grammatical. (In a more informal experiment, we have found that colleagues who read papers on GPs, understand new GPs easily by the end of a paper.) LAZY exhibits this behavior by being more careful after encountering SDGPs or when reanalyzing garden path sentences.

Current Address:
The Aerospace Corporation
P.O. Box 92957
Los Angeles, CA 90009

II. SYNTAX IN A CONCEPTUAL ANALYZER

The goal of conceptual analysis is to map natural language text into memory structures that represent the meaning of the text. It is claimed that this mapping can be accomplished without a prior syntactic analysis, relying instead on a variety of knowledge sources including expectations from both word definitions and inferential memory (see [Riesbeck 76], [Schank 80], [Gershman 82], [Birnbaum 81], [Pazzani 83] and [Dyer 83]). Given this model of processing, in sentence (4),

(4) Mary kicked John.

How is it possible to tell who kicked whom? There is a very simple answer: Syntax. Sentence (4) is a simple active sentence whose verb is "to kick". "Mary" is the subject of the sentence and "Bill" is the direct object. There may be a more complicated answer, if, for example, John and Mary are married, Mary is ill-tempered, John is passive, and Mary has just found out that John has been unfaithful. In this case, it is possible to expect that Mary might hit John, and confirm this prediction by noticing that the words in (4) refer to Mary, John, and hitting. In fact, if this prediction was formulated and the sentence were "John kicked Mary" we might take it to mean "Mary kicked John" and usually notice that the speaker had made a mistake. Although we feel that this type of processing is an important part of understanding, it cannot account for all language comprehension. Certainly, (4) can be understood in contexts which do not predict that Mary might hit John, requiring syntactic knowledge to determine who kicked whom.

IIa. Precedes and Follows

Syntactic information is represented in a conceptual analyzer, in a number of ways, the simplest of which is the notion of one word preceding or following another. Such information is encoded as a positional predicate in the test of a type of production which Riesbeck calls a request. The test also contains a semantic predicate (i.e., the selectional restrictions). A set of requests make up the definition of a word. For example, the definition of "kick" has three requests:

```
REQ1:   Test:     true
        Action:   Add the meaning structure
                  for "kick" to an ordered
                  list of concepts typically
                  called the C-list.

REQ2:   Test:     Is there a concept
                  preceding the concept for
                  "kick" which is animate?
        Action:   ...

REQ3:   Test:     Is there a concept
                  following the concept for
                  "kick" which is a physical object?
        Action:   ...
```

The action of a request typically builds or connects concepts. Although people who build conceptual analyzers have reasons for not building a representation of the syntax of a sentence, there is no reason that they can not. LAZY builds syntactic representations.

IIb. Requests in LAZY

LAZY, unlike other conceptual analyzers, separates the syntactic (or positional) information from the selectional restrictions by dividing the test part of request into a number of facets. There are three reasons for doing this. First, it allows for a distinction between different kinds of knowledge. Secondly, it is possible to selectively ignore some facets. Finally, it permits a request to access the information encoded in other requests.

In many conceptual analyzers, some syntactic information is hidden in the control structure. At certain times during the parse, not all of the request are considered. For example, in (5) it is necessary to delay considering a request.

(5) Who is Mary recruiting?

To avoid understanding the first three words of sentence (5) as a complete sentence, "Who is Mary?", some request from "is" must be delayed until the word "recruiting" is processed. In LAZY, the time that a request can be considered is explicitly represented as a facet of the request. Additionally, separate tests exist for the selectional restriction, the expected part of speech, and the expected sentential position.

In LAZY, REQ2 of "kick" would be:

```
REQ2a:  Position: Subject of "kick"
        Restriction: Animate
        Action:       Make the concept
                      found the syntactic
                      subject of "kick"
        Part-Of-Speech: (noun pronoun)
        Time: Clause-Type-Known?
```

In REQ2a, Subject is a function which examines the state of the C-list and returns the proper constituent as a function of the clause type. In an active declarative sentence, the subject precedes the verb, in a passive sentence it may follow the word "by", etc. (The usage of "subject" is incorrect in the usual sense of the word.) The Time facet of REQ2a states that the request should be considered only after the type of the clause is know. The predicates which are included in a request to control the time of consideration are: End-Of-Noun-Group?, Clause-Type-Known?, Head-Of-Immediate-Noun-Group?, and End-Of-Sentence?. These operate by examining the C-list in a manner similar to the positional predicates. The other facets of REQ2a state that the subject of "kick" must be animate, and should be a noun or a pronoun.

III GARDEN PATH SENTENCES

Several different types of local ambiguities cause GPs. Misunderstanding sentences 1, 2 and 3 is a result of confusing a participle for the main verb of a sentence. Although there are other types of GPs (e.g., imperative and yes/no questions with an initial "have"), we will only demonstrate how LAZY understands or misunderstands passive participle and main verb conflicts.

Passive participles and past main verbs are indicated by a "ed" suffix on the verb form. Therefore, the definition of "ed" must discriminate between these two cases. The definition of "ed" is shown in Figure 3a. A simpler definition for "ed" is possible if the morphology routine reconstructs sentences so that the suffix of a verb is a separate "word" which precedes the verb. The definition of "ed" is shown in Figure 3a. Throughout this discussion, we will use the name Root for the verb immediately following "ed" on the C-list.

If Root appears to be passive
Then mark Root as a passive participle.
Otherwise if Root does not appear to be passive
Then note the tense of Root.

Figure 3a. Definition of "ed".

It is safe to consider this request only at the end of the sentence or if a verb is seen following Root which could be the main verb. One test that is used to determine if Root could be passive is:

1. There is no known main verb seen preceding "ed", and
2. The word which would be the subject of Root if Root were active agrees with the selectional restrictions for the word which would precede Root if Root were passive (i.e., the selectional restrictions of the direct object if there is no indirect object), and
3. There is a verb which could be the main verb following Root.

Figure 3b.

One test performed to determine if Root does not appear to be passive is:

1. The verb is not marked as passive, and
2. The word which would be the subject of Root if Root were active agrees with the selectional restrictions for the subject.

Figure 3c.

Note that these tests rely on the fact that one request can examine the semantic or syntactic information encoded in another request.

As we have presented requests so far, four separate tests must be true to fire a request (i.e., to execute the request's action): a word must be found in a particular position in the sentence, the word must have the proper part of speech, the word must meet the selectional restrictions, and the parse must be in a state in which it is safe to execute the positional predicate. We have relaxed the requirement that the selectional restrictions be met if all of the other tests are true. This avoids problems present in some previous conceptual analyzers which are unable to parse some sentences such as "Do rocks talk?". Additionally, we have experimented with not requiring that the Time test succeed if all other tests have passed unless we are reanalyzing a sentence that we have previously not been able to parse. We will demonstrate that this yields the performance that people exhibit when comprehending GPs.

LAZY processes a sentence one word at a time from left to right. When processing a word, its representation is added to the C-list and its requests are activated. Next, all active requests are considered. When a request is fired, a syntactic structure is built by connecting two or more constituents on the C-list. At the end of a parse the C-list should contain one constituent as the root of a tree describing the structure of the sentence.

Sentence (6) is a GP which people normally have trouble reading:

(6) The boat sailed across the river sank.

When parsing this sentence, LAZY reads the word "the" and adds it to the C-list. Next, the word "boat" is added to the C-list. A request from "the" looking for a noun to modify is considered and all tests pass. This request constructs a noun phrase with "the" modifying "boat". Next, "ed" is added to the C-list. All of its requests look for a verb following, so they can not fire yet. The work "sail" is added to the C-list. The request of "ed" which sets the tense of the immediately following verb is considered. It check the semantic features of "boat" and finds that they match the selectional restrictions required of the subject of "sail". The action of this request is executed, in spite of the fact that its Time reports that it is not safe to do so. Next, a request from "sail" finds that that "boat" could serve as the subject since it precedes the verb in what is erroneously assumed to be an active clause. The structure built by this request notes that "boat" is the subject of "sail". A request looking for the direct object of "sail" is then considered. It notices that the subject has been found and it is not animate, therefore "sail" is not being used transitively. This request is deactivated. The word "across" is added to the C-list and "the river" is then parsed analogously to "the boat". Next, a request from "across" looking for the object of the preposition is considered and finds the noun phrase, "the river". Another request is then activated and attaches this prepositional phrase to "sail". At this point in the parse, we have built a structure describing an active sentence "The boat sailed across the river." and the C-list contains one constituent. After adding the verb suffix and "sink" to the C-list we find that "sink" cannot find a subject and there are two constituents left on the C-list. This is an error condition and the sentence must be reanalyzed more carefully.

It is possible to recover from misreading some garden path sentences by reading more carefully. In LAZY, this corresponds to not letting a request fire until all the tests are true. Although other recovery schemes are possible, our current implementation starts over from the beginning. When reanalyzing (6), the request from "ed" which sets the tense of the main verb is not fired because all facets of its test never become true. This request is deactivated when the word "sank" is read and another request from "ed" notes that "sailed" is a participle. At the end of the parse there is one constituent left on the C-list, similar to that which would be produced when processing "The boat which was sailed across the river sank".

It is possible to parse SDGPs without reanalysis. For example, most readers easily understand (7) which is simplified from [Birnbaum 81].

(7) The plane stuffed with marijuana crashed.

Sentence (7) is parsed analogously to (6) until the word "stuff" is encountered. A request from "ed" tries to determine the sentence type by testing if "plane" could be the subject of "stuff" and fails because "plane" does not meet the selectional restrictions of "stuff". This request also checks to see if "stuff" could be passive, but fails at this time (see condition 3 of Figure 3b). A request from "stuff" then finds that "plane" is in the default position to be the subject, but its action is not executed because two of the four tests have not passed: the selectional restrictions are violated and it is too early to consider the positional predicate because the sentence type is unknow. A request looking for the direct object of "stuff" does not succeed at this time because the default location of the direct object follows the verb. Next, the prepositional phrase "with marijuana" is parsed analogously to "across the lake" in (6). After the suffix of "crash" (i.e., "ed") and "crash" are added to the C-list, the request from the "ed" of "stuff" is considered, and it finds that "stuff" could be a passive participle because "plane" can fulfill the selectional restrictions of the direct object of "stuff". A request from "stuff" then notes that "plane" is the direct object, and a request from the "ed" of "crash" marks the tense of "crash". Finally, "crash" finds "plane" as its subject. The only constituent of the C-list is a tree similar to that which would be produced by "The plane which was stuffed with marijuana crashed".

There are some situations in which garden path sentences cannot be understood even with a careful reanalysis. For example, many people have problems understanding sentence (8).

(8) The canoe floated down the river sank.

To help some people understand this sentence, it is necessary to inform them that "float" can be a transitive verb by giving a simple example sentence such as "The man floated the canoe". Our parser would fail to reanalyze this sentence if it did not have a request associated with "float" which looks for a direct object.

We have been rather conservative in giving rules to determine when "ed" indicates a past participle instead of the past tense. In particular, condition 3 of Figure 3b may not be necessary. By removing it, as soon as "the plane stuffed" is processed we would assume that "stuffed" is a participle phrase. This would not change the parse of (7). However, there would be an impact when parsing (9).

(9) The chicken cooked with broccoli.

With condition 3 removed, this parses as a noun phrase. With it included, (9) would currently be recognized as a sentence. We have decided to include condition 3, because it delays the resolving of this ambiguity until both possibilities are clear. It is our belief that this ambiguity should be resolved by appealing to episodic and conceptual knowledge more powerful than selectional restrictions.

IV. PREVIOUS WORK

In PARSIFAL, Marcus' parser, the misunderstanding of GPs is caused by having grammar rules which can look ahead only three constituents. To deterministically parse a GP such as (1), it is necessary to have a look ahead buffer of at least four constituents. PARSIFAL's grammar rules make the same guess that readers make when presented with a true GP. For a participle/main verb conflict, readers prefer to choose a main verb. However, PARSIFAL will make the same guess when processing SDGPs. Therefore, PARSIFAL fails to parse some sentences (SDGPs) deterministically which people can parse without conscious backtracking. In LAZY, the C-list corresponds to the look ahead buffer. When parsing most sentences, the C-list will contain at most three constituents. However, when understanding a SDGP or reanalyzing a true garden path sentence, there are four constituents in the C-list. Instead of modeling the misunderstanding of GPs, by limiting the size of the look-ahead buffer and the look ahead in the grammar, LAZY models this phenomenon by deciding on a syntactic representation before waiting long enough to disambiguate on a purely syntactic basis when semantic expectations are strong enough.

Shieber models the misunderstanding of GPs in a LALR(1) parser [Aho 77] by the selection of an incorrect reduction in a reduce-reduce conflict. In a participle/main verb conflict, there is a state in his parser which requires choosing between a participle phrase and a verb phrase. Instead of guessing like PARSIFAL, Shieber's parser looks up the "lexical preference" of the verb. Some verbs are marked as preferring participle forms; others prefer being main verbs. While this lexical preference can account for the understanding of SDGPs and the misunderstanding of GPs in any one particular example, it is not a very general mechanism. One implication of using lexical preference to select the correct form is that some verbs are only understood or misunderstood as main verbs and others only as participles. If this were true, then sentences (10a) and (10b) would both be either easily understood or GPs.

(10a) No freshmen registered for Calculus failed.

(10b) No car registered in California should be driven in Mexico.

We find that most people easily understand (10b), but require conscious backtracking to understand (10a). Instead of using a predetermined preference for one syntactic form, LAZY utilizes semantic clues to favor a particular parse.

V. FUTURE WORK

We intend to extend LAZY by allowing it to consult and episodic memory during parsing. The format that we have chosen for requests can be augmented by adding an EPISODIC facet to the test. This will enable expectation to predict individual objects in addition to semantic features. We have seen examples of potential garden path sentences which we speculate are misunderstood or understood by consulting world knowledge (e.g., 11 and 12)

(11) At MIT, ninety five percent of the freshmen registered for Calculus passed.

(12) At MIT, five percent of the freshmen registered for Calculus failed.

We have observed that more people mistake "registered" for the main verb in (11) than (12). This could be accounted forby the fact that the proposition that "At MIT, ninety five percent of the freshmen registered for Calculus" is more easily accepted than "At MIT, five percent of the freshmen registered for Calculus". Evidence such as this suggests that semantic and episodic processing are done at early stages of understanding.

VI. CONCLUSION

We have augmented the basic request consideration algorithm of a conceptual analyzer to include information to determine the time that an expectation should be considered and shown that by ignoring this information when syntactic and semantic expectations agree, we can model the performance of native English speakers understanding and misunderstanding garden path sentences.

VII. ACKNOWLEDGMENTS

This work was supported by USAF Electronics System Division under Air Force contract F19628-84-C-0001 and monitored by the Rome Air Development Center.

BIBLIOGRAPHY

Birnbaum, L. and M. Selfridge, "Conceptual Analysis of Natural Language", in Inside Artificial Intelligence: Five Programs Plus Miniatures, Hillsdale, NJ: Lawrence Erlbaum Associates, 1981.

Crain, S. and P. Coker, "A Semantic Constraint on Parsing", Paper presented at Linguistic Society of America Annual Meeting. University of California at Irvine, 1979.

Dyer, M.G., In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension, Cambridge, MA: The MIT Press, 1983.

Gershman, A.V., "A Framework for Conceptual Analyzers", in Strategies for Natural Language Processing, Hillsdale, NJ: Lawrence Erlbaum Associates, 1982.

Katz, J. S. and J. A. Fodor, "The Structure of Semantic Theory", in Language, 39, 1963.

Marcus, M., A Theory of Syntactic Recognition for Natural Language, Cambridge, MA: The MIT Press, 1980.

Marcus, M., "Wait-and-See Strategies for Parsing Natural Language", MIT WP-75, Cambridge, MA: 1974.

Matthews, R., "Are the Grammatical Sentences of a Language of Recursive Set?", in Systhese 40, 1979.

Pazzani, M. J., "Interactive Script Instantiation", in Proceedings of the National Conference on Artificial Intelligence, 1983.

Riesbeck, C. and R. C. Schank, "Comprehension by Computer: Expectation Based Analysis of Sentences in Context", Research Report #78, Dept. of Computer Science, Yale University, 1976.

Schank, R. C. and L. Birnbaum, Memory, Meaning, and Syntax, Research Report 189, Yale University Department of Computer Science, 1980.

Shieber, S. M., "Sentence Disambiguation by a Shift-Reduce Parsing Technique", 21st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 1983.