Parsing

W. A. Martin
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Looking at the Proceedings of last year's Annual Meeting, one sees that the session most closely paralleling this one was entitled Language Structure and Parsing. In a very nice presentation, Martin Kay was able to unite the papers of that session under a single theme. As he stated it,

> There has been a shift of emphasis away from highly structured systems of complex rules as the principal repository of information about the syntax of a language towards a view in which the responsibility is distributed among the lexicon, semantic parts of the linguistic description, and a cognitive or strategic component. Concomitantly, interest has shifted from algorithms for syntactic analysis and generation, in which the central structure and the exact sequence of events are paramount, to systems in which a heavier burden is carried by the data structure and in which the order of events is a matter of strategy.

This year, the papers of the session represent a greater diversity of research directions. The paper by Hayes, and the paper by Wilensky and Arens are both examples of what Kay had in mind, but the paper by Church, with regard to the question of algorithms, is quite the opposite. He holds that once the full range of constraints describing people's processing behavior has been captured, the best parsing strategies will be rather straightforward, and easily explained as algorithms.

Perhaps the seven papers in this year's session can best be introduced by briefly citing some of the achievements and problems reported in the works they reference.

In the late 1960's Woods [Woods70] capped an effort by several people to develop ATN parsing. This well known technique applies a straightforward top down, left to right, depth first parsing algorithm to a syntactic grammar. Especially in the compiled form produced by Burton [Burton76a], the parser was able to produce the first parse in good time, but without semantic constraints, numerous syntactic analyses could be and sometimes were found, especially in sentences with conjunctions. A strength of the system was the ATN grammar, which can be described as a set of context free production rules whose right hand sides are finite state machines and whose transition arcs have been augmented with functions able to read and set registers, and also able to block a transition on their arc. Many people have found this a convenient formulism in which to develop grammars of English.

The Woods ATN parser was a great success, and attempts were made to exploit it (a) as a model of human processing and (b) as a tool for writing grammars. At the same time it was recognized to have limitations. It wasn't tolerant of errors, and it couldn't handle unknown words or constructions (there were many syntactic constructions which it didn't know). In addition, the question answering system fed by the parser had a weak notion of word and phrase semantics and it was not always able to handle quantifiers properly. It is not clear these components could have supported a stronger interaction with syntactic parsing, had Woods chosen to attempt it.

On the success side, Kaplan [Kaplan72] was inspired to claim that the ATN parser provided a good model for some aspects of human processing. Some aspects which might be modeled are:

Linguistic Phenomenon	ATN Computational Mechanism
Preferred readings of Ambiguous Sentences	Ordered Trying of Alternative Arcs
Garden Path Sentences	Back-tracking
Perceived Complexity Differences	Hold List Costing Counting Total Transitions
Center Embedding Bounds	None

In one study, most people got the a) reading of 1). One can try to explain this

1) They told the girl that Bill liked the story.

1a) They told the girl [that [Bill liked the story]$_S$].

1b) They told [the girl that Bill liked]$_{NP}$ the story.

by ordering the arcs leaving the state where the head noun of an NP has been accepted: a pop arc (terminating the NP) is tried before an arc accepting a modifying relative clause. However, Rich [Rich75] points out that this arc ordering solution would seem to have difficulties with 2). This sentence is often not perceived

2) They told the girl that Bill liked that he would be at the football game.

as requiring backup, yet if the arcs are ordered as for 1), it does require backup. There is no doubt that whatever is going on, the awareness of backup in 3) is so much stronger than in 2) that it seems like a different phenomenon. To resolve this,

3) The horse raced past the barn fell.

one could claim that perceived backup is some function of the length of the actual backup, or maybe of the degree of commitment to the original path (although it isn't clear what this would mean in ATN terms).

In this session, Ferrari and Stock will turn the arc ordering game around and describe, for actual texts, the probability that a given arc is the correct exit arc from a node, given the arc by which the parser arrived at the node. It will be interesting to look at their distributions. In the speech project at IBM Watson Laboratories [Baker75] it was discovered some time ago that, for a given text, the syntactic class of a word could be predicted correctly over 90% of the time given only the syntactic class of the preceding word. Interestingly, the correctness of predictions fell off less than 10% when only the current word was used. One wonders if this same level of skewness holds across texts, or (what we will hear) for the continuation of phrases. These results should be helpful in discussing the whole issue of arc ordering.

Implicit in any arc ordering strategy is the assumption that not all parses of a sentence will be found. Having the "best" path, the parser will stop when it gets an acceptable analysis. Arc ordering helps find that "best" path. Marcus [Marcus78], agreed with the idea of following only a best path, but he claimed that the reason there is no perceived backup in 2) is that the human parser is able to look ahead a few constituents instead of just one state and one constituent in making a transition. He claims this makes a more accurate model of human garden path behavior, but it doesn't address the issue of unlimited stack depth. Here, Church will describe a parser similar in design to Marcus', except that it conserves memory. This allows Church to address psychological facts not addressed by either Marcus or the ATN models. Church claims that exploiting stack size constraints will increase the chances of building a good best path parser.

Besides psychological modeling, there is also an interest in using the ATN formalism for writing and teaching grammars. Paramount here is explanation, both of the grammar and its application to a particular sentence. The paper by Kehler and Woods reports on this. Weischedel picks a particular problem, responding to an input which the ATN can't handle. He associates a list of diagnostic conditions and actions with each state. When no parse is found, the parser finds the last state on the path which progressed the farthest through the input string and executes its diagnostic conditions and actions. When a parser uses only syntactic constraints, one expects it to find a lot of parses. Usually the number of parses grows more than linearly with sentence length. Thus, for a fairly complete grammar and moderate to long sentences, one would expect that the case of no parses (handled by Weischedel) would be rare in comparison with the other two cases (not handled) where the set of parses doesn't include the correct one, or where the grammar has been mistakenly written to allow undesired parses. Success of the above efforts to follow only the best path would clearly be relevant here. No doubt Weischedel's procedure can help find a lot of bugs if the test examples are chosen with a little care. But there is still interesting work to be done on grammar and parser explanation, and Weischedel is one of those who intends to explore it.

The remaining three papers stem from three separate traditions which reject the strict syntactic ATN formalism, each for its own reasons. They are:

i) Semantic Grammars -- the Davidson and Kaplan paper

ii) Semantic Structure Driven Parsing -- Wilensky and Arens paper

iii) Multiple knowledge Source Parsing -- Hayes paper

Each of these systems claims some advantage over the more widely known and accepted ATN.

The semantic grammar parser can be viewed as a variation of the ATN which attempts to cope with the ATN's lack of semantics. Kaplan's work builds on work started by Burton [Burton76b] and picked up by Hendrix et al [Hendrix78]. The semantic grammar parser uses semantic instead of syntactic arc categories. This collapses syntax and semantics into a single structure. When an ATN parsing strategy is used the result is actually less flexible than a syntactic ATN, but it is faster because syntactic possibilities are eliminated by the semantics of the domain. The strategy is justified in terms of the performance of actual running systems. Kaplan also calls on a speed criteria in suggesting that when an unknown word is encountered the system assume all possibilities which will let parsing proceed. Then if more than one possibility leads to a successful parse, the system should attempt to resolve the word further by file search or user query.

As Kaplan points out, this trick is not limited to semantic grammars, but only to systems having enough constraints. It would be interesting to know how well it would work for systems using Osherson's [Osherson78] predicability criterion, instead of truth for their semantics. Osherson distinguishes between "green idea", which he says is silly and "married bachelor" which he says is just false. He notes that "idea is not green" is no better, but "bachelor is not married" is fine. Predicability is a looser constraint than Kaplan uses, and if it would still be enough to limit database search this would be interesting, because predicability is easier to implement across a broad domain.

Wilensky is a former student of Schank's and thus comes from a tradition which emphasizes semantics over syntax. He is right in emphasizing the importance of phrase semantics. The grammarians Quirk and Greenbaum [Quirk73] point out the syntactic and semantic importance of verb phrases over verbs. In linguistics, Bresnan [Bresnan80] is developing a theory of lexical phrases which accounts, by lexical relations between constituents of a phrase, for many of the phenomena explained by the old transformational grammar. For example, given

4) There were reported to have been lions sighted.

a typical ATN parser would attempt by register manipulations to make "lions" the subject. Using a phrase approach, "there be lions sighted" can be taken as meaning "exist lions sighted," where "lions" is an object and "sighted" an object complement. "There" is related to the "be" in "been" by a series of relationships between the arguments of semantic structures. Wilensky appears to have suppressed syntax into his semantic component, and so it will be interesting to see how he handles the traditional syntactic phenomena of 4), like passive and verb forms.

Finally, the paper by Hayes shows the influence of the speech recognition projects where bad input gave the Woods ATN great difficulty. Text input is much better than speech input. However, examination of actual input [Malhotra75] does show sentences like:

5) What would have profits have been?

Fortunately, these cases are rare. Much more likely is elipsis and the omission of syntax when the semantics are clear. For example, the missing commas in

6) Give ratios of manufacturing costs to sales for plants 1 2 3 and 4 for 72 and 73.

Examples like these show that errors and omissions are not random phenomena and that there can be something to the study of errors and how to deal with them.

In summary, it can be seen that while much progress has been made in constructing usable parsers, the basic issues, such as the division of syntax, semantics, and pragmatics both in representation and in order of processing, are still up for grabs. The problem has plenty of structure, so there is good fun to be had.

References

[Baker75] Baker, J.K. "Stochastic Modeling for Automatic Speech Understanding," Speech Recognition: Invited Papers of the IEEE Symposium, Reddy, D.R. (Ed.), 1975.

[Bresnan80] Bresnan, Joan. "Polyadicity: Part I of a Theory of Lexical Rules and Representations," MIT Department of Linguistics (January 1980).

[Burton76a] Burton, Richard R. and Woods, William A. "A Compiling System for Augmented Transition Networks," COLING 76.

[Burton76b] Burton, Richard R. "Semantic Grammar: An Engineering Technique for Constructing Natural Language Understanding Systems," BBN Report 3453, Bolt, Beranek, and Newman, Boston, Ma. (December 1976).

[Hendrix78] Hendrix, Gary G., Sacerdoti, E.D., Sagalowicz, D., and Slocum, J. "Developing a Natural Language Interface to Complex Data," ACM Trans. on Database Systems, vol. 3, no. 2 (June 1978), pp. 105-147.

[Kaplan72] Kaplan, Ronald M. "Augmented Transition Networks as Psychological Models of Sentence Comprehension," Artificial Intelligence, 3 (October 1972), pp. 77-100.

[Malhotra75] Malhotra, Ashok. "Design Criteria for a Knowledge-Based English Language System for Management: An Experimental Analysis," MIT/LCS/TR-146, MIT, Laboratory for Computer Science, Cambridge, Ma. (February 1975).

[Marcus78] Marcus, Mitchell. "A Theory of Syntactic Recognition for Natural Languages," Ph.D. thesis, MIT Dept. of Electrical Engineering and Computer Science, Cambridge, Ma. (to be published by MIT Press).

[Osherson78] Osherson, Daniel N. "Three Conditions on Conceptual Naturalness," Cognition, 6 (1978), pp. 263-289.

[Quirk73] Quirk, R. and Greenbaum, S. A Concise Grammar of Contemporary English, Harcourt Brace Jovanovich, New York (1973).

[Rich75] Rich, Charles. "On the Psychological Reality of Augmented Transition Network Models of Sentence Comprehension," unpublished paper, MIT Artificial Intelligence Laboratory, Cambridge, Ma. (July 1975).

[Woods70] Woods, William A. "Transition Network Grammars for Natural Language Analysis" CACM 13, 10 (October 1970), pp. 591-602.