Reconnaissance-Attack Parsing*

Michael B. Kac
Department of Linguistics
University of Minnesota
Minneapolis, MN 55455 USA

Tom Rindflesch
Department of Linguistics
University of Minnesota
Minneapolis, MN 55455 USA

Karen L. Ryan
Computer Sciences Center
Honeywell, Inc.
Minneapolis, MN 55427 USA

In this paper we will describe an approach to parsing one major component of which is a stragegy called RECONNAISSANCE-ATTACK. Under this strategy, no structure building is attempted until after completion of a preliminary phase designed to exploit low-level information to the fullest possible extent. This first pass then defines a set of constraints that restrict the set of available options when structure building proper begins. R-A parsing is in principle compatible with a variety of different views regarding the nature of syntactic representation, though it fits more comfortably with some than with others--a point to which we shall return.

Three principles which are closely allied with Reconnaissance-Attack are MULTIPLE OPTIONS, GRADED ATTACK and ALTRUISM AVOIDANCE. The principle of Multiple Options states that a variety of techniques for solving particular problems should be available, ranging from simple and constrained to complex and powerful. The underlying idea is that while it might be demonstrable that certain tasks lie beyond the capabilities of, say, a local pattern matcher, the pattern matcher might suffice often enough to warrant its retention for the cases where it is effective. The corollary principle of Graded Attack states that more powerful weapons should be called up only after less powerful ones have failed. Finally, the principle of Altruism Avoidance states (in its strongest form) that no computational effort should be undertaken without the guarantee of a payoff. Limitations of space preclude a detailed treatment of the first two here (see Rindflesch forthcoming), though the third is implicit in the approach to specific examples to be mentioned later on.

The extent to which these ideas are valid may vary across applications. In psychological modelling it may prove undesirable to exploit them fully given the existence of phenomena ('garden path effects') which suggest that in human linguistic processing, attack is sometimes premature or based on faulty reconnaissance.[1] On the other hand, the principles we have described have attractive properties in the context of purely practical goals.

It is widely agreed that a satisfactory approach to parsing must minimize backtracking, storage of uneliminated options and parallel consideration of alternatives; R-A parsing is directly addressed to this problem. (How it differs from other approaches should soon become clear.) Given a question regarding the structure of a sentence, the only possible responses by a pure R-A parser are 'right answer' and 'no answer'. That is, if at a given point there is an indeterminacy as to how to proceed, the currently active program module simply 'passes'. (Certain ambiguities are thus treated as involving situations where, given several available options, no resolution has been made after all relevant information has been taken into account.) As many decisions as can be reliably made at a given stage of analysis actually are made, but any which cannot be made at that point are deferred to later stages. The kinds of decisions made during the initial stages of parsing do not constitute structure building (i.e. construction of a parse tree or formally equivalent object), though they constrain the range of options available when structure building actually begins.

The rationale for reconnaissance is the fact that the resolution of local indeterminacies can depend on following material, and often does. Working strictly left-right, even with a 'predictive' parser, obviously runs afoul of this problem. Lookahead (Marcus 1980) avoids the difficulty to some extent, but could be regarded as linguistically unmotivated. That is, the need to invoke it in the first place is merely an artifact of failure to capitalize on global syntactic information which can be easily recovered from the input before any attempt at structure building is made. As a simple example, consider a sentence like *I believe you* as opposed to e.g. *I believe you did it.* If reconnaissance first counts the number of verbs in the sentence, and finds only one, then the possibility that *you* in the first example could initiate a subordinate clause can be eliminated. In the second case, that the Object of *believe* is not an NP can also be determined from cues made available by reconnaissance; see Rindflesch forthcoming.

A particularly troublesome problem to which an R-A-based solution seems ideally suited is the ubiquitous category-label ambiguity (CLA) found in English. While some instances of local CLA can be resolved on the basis of preceding material, this cannot be guaranteed; indeed, ambiguities can pile up awaiting disambiguation by later material; as an example, consider the CLA exhibited by the first three words in *Leave time forms the basis for faculty renewal*. (Given just these words, there are at least five distinct structural configurations which could be initiated by this sequence.) In the approach described in Rindflesch op. cit., it is determined during reconnaissance that it is not possible for more than one of the first three words to be a verb owing to the impossibility of constructing a legal 'ordination configuration' (a representation of the relations of sub- and superordination that obtain among predicates in the sentence.) This follows from the fact that none of the potential verbs can take complements in either the Subject or Object relations, and that necessary conditions for the occurrence of other kinds of predicate-containing arguments (such as the presence of overt subordinators, like relative pronouns) are not satisfied. It is also possible, PRIOR TO ANY STRUCTURE BUILDING, to determine that *forms* is the actual verb, though limitations of space prohibit a detailed explanation. (Again, see Rindflesch op. cit.) This approach contrasts vividly with more traditional ones, which typically use such brute-force methods as backtracking to revoke incorrect hypotheses or parallel consideration (actual or simulated) of uneliminated alternatives.[2]

This criticism might seem misplaced since it is an accepted practice for grammars which drive parsers to attach a 'likelihood' estimate to each rule to aid in the determination of which rule to apply next during a parse. But such values are ad hoc assignments which do not further our theoretical understanding of parsing natural language; a blind guess supported by a likelihood estimate may have a statistically better chance of success, but is a blind guess nonetheless. (Some authors recognize this fact, e.g. Proudian and Pollard (1985), and are careful to separate the ad hoc heuristic assignment of likelihood measures from the operation of the parser in general.) Having a reconnaissance phase decreases the extent to which it is necessary to rely on blind application of grammar rules. The essential idea is to exploit to the fullest possible extent knowledge about the parsing task derived from the particular properties of the terrain, in contrast to approaches which use heuristic likelihood estimates as an add-on to the actual grammar or parser. (For further discussion, see Ryan and Rindflesch 1986.)

The theoretical framework assumed in our work on R-A parsing is corepresentational grammar (Kac 1978, 1985), which has the advantage of facilitating precise manipulation of traditional grammatical notions such as logical Subject and Object, and sub- and superordination. These traditional notions are important components in the formulation of linguistically insightful strategies for parsing, as illustrated by the critical role played by 'ordination' relations in the preceding paragraph.

In order to implement this sort of strategy, it is necesary to impose some additional control on the parse beyond what would typically be required in a less structured approach. Though multiple passes over the input are required and it is necessary to explicitly maintain several different sources of information beyond the traditional parse tree, the payoff for this additional control and data is the ability to bypass redundant intermediate stages in the parse when the input contains sufficient information to allow it. Parsing is thus highly data-driven in the sense that only as much machinery as is absolutely necessary is used for any particular parse, and the specific devices used will vary from case to case. The cost of multiple passes, moreover, can be kept relatively low given that the goal of each pass is so specialized as to assure that it can be completed quickly, while reconnaissance guarantees that the absolute number of passes required, even in very complex examples, can be kept to a minimum.

It is helpful to contrast what we have in mind to a system of phrase structure rules. Such a system compresses together a variety of different types of information (e.g. information about dependencies, categories, grammatical relations and subcategorization), and this compression makes it difficult to isolate just the subset of these information types which is most relevant to the problem while excluding information which is redundant or irrelevant to the current decision point. It seems to us less costly and more revealing to organize the grammar underlying the parser in a more 'atomistic' way, thus making it unnecessary to tease out of the rules information which they do not encode in a transparent fashion. For example, the rules

S -> NP VP   NP -> ART N   VP -> V S

(if taken by themselves) imply that if a sentence contains two or more verbs, each noninitial one must be in its own embedded clause. From this it can be deduced that there can be only one main verb, but this information (which figures crucially in the definition of a legal ordination configuration) is not represented in an immediately accessible form. An added complication is represented by the fact that while arbitrary proper subsets of a set of PS-rules indicate what is PERMITTED, there is no way except from consideration of the grammar as a whole to determine what is PROHIBITED.

Our approach facilitates the use of information at stages in the parsing process where that information is most useful, which has consequences not just for syntactic parsing alone but in integrating the syntactic and semantic aspects of the understanding process. It has long been recognized that some semantic decisions can be made before an entire syntactic parse is available, and that the results of these semantic decisions can be used to drive further results in the syntactic parse. It is further recognized that such early semantic processing makes good sense computationally, and some current systems make good use of this principle. We take this idea one step further by allowing the parser to anticipate on the basis of very rudimentary and low level cues many structural characteristics of the input which traditional approaches cannot recognize until substantial structure building has been done. For example, in the R-A model described in Kac 1981, it is possible to delineate the boundaries of complex NP's before their precise constituency is known.

As noted earlier, R-A parsing is in principle consistent with a variety of assumptions about syntactic representation, and is not rigidly tied to the assumption that the end result of syntactic parsing is a traditional phrase structure tree. (See Kac and Manaster-Ramer 1986 for discussion.) The goal of parsing is to provide an input to the semantic component, and such an input can in principle take a variety of forms (such as a representation of predicate-argument structure).

It is worth pointing out, if it is not already obvious, that we see the central issues as linguistic ones first and computational ones second in the sense that the kind of approach which seems to us to hold out the most promise is one in which efficient parsing is the product largely of an adequate qualitative picture of linguistic structure. This picture tells us two crucial things: what information there is in the input to be exploited, and when it first becomes available. We hope to have given at least a preliminary indication of how such a picture can contribute to insightful solutions to interesting problems in natural language processing.

Notes

*The listing of authors is strictly alphabetical.

1. On the other hand, a compromise model in which some features of R-A parsing are exploited has some attractions. For example, suppression of the optional *that*-complementizer in sentences like *I believe (that) Mary likes Bill* slightly increases comprehension difficulty, a phenomenon which can be naturally interpreted as a short-lived garden path. One possible way to distinguish between effects such as the one just mentioned and garden pathing from which there is evidently no possibility of recovery is to allow some guess-and-back-up processing in the reconaissance phase and to attribute short-lived effects to garden pathing prior to the onset of the structure building (attack) phase.

2. We assume that lexical lookup for the entire sentence is done before any syntactic processing takes place. This yields the advantage of increased modularity, as compared to a system in which lexical lookup is incorporated into the syntax; it is thus possible to make modifications and revisions in the part of the parser which deals with CLAR without the need for corresponding revisions in other modules.

References

Kac, M.B. 1978. Corepresentation of Grammatical Structure. Minneapolis and London: University of Minnesota Press and Croom-Helm.

---. 1981. Center-embedding revisited. Proceedings of the Third Annual Meeting of the Cognitive Science Society.

---. 1985. Grammars and Grammaticality. Unpublished ms., University of Minnesota.

--- and A. Manaster-Ramer. 1986. Parsing without (much) constituent structure. In this volume.

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

Proudian, D. and C. Pollard. 1985. Parsing head-driven phrase structure grammar. Proceedings of the 23rd Annual Meeting of the Association for Compuational Linguistics.

Rindflesch, T. forthcoming. Doctoral Dissertation in preparation, University of Minnesota.

--- and K.L. Ryan. 1986. Resolution of category label ambiguity by Reconnaissance-Attack Parsing. Unpublished paper, University of Minnesota.

Ryan, K.L. and T. Rindflesch. 1986. A Theory of Heuristics for Parsing. Submitted to AAAI Annual Meeting.