A NATIONAL RESOURCE GRAMMAR

Jerry R. Hobbs

Artificial Intelligence Center
SRI International
Menlo Park, California 94025

1. THE PROBLEM AND ITS SOLUTION

The syntax of English is largely a solved problem. Yet all natural language projects devote a large amount of their effort to developing grammars. The reason for this situation is that there is no very large, generally available grammar of English based on current technology—unification grammar. The solution is to develop a very broad-coverage National Resource Grammar in a unification formalism, perhaps under the auspices of the Linguistic Data Consortium (LDC) and freely available to its members.

What do we mean when we say syntax is a solved problem? The syntactic structure and the corresponding predicate-argument, or operator-operand, relations are worked out for a great majority of grammatical constructions. Moreover, they are largely agreed upon, modulo some fairly easily resolvable theoretical differences in representation. Syntax is still a healthy area of research, but most of the work is concentrated on achieving more elegant treatments and characterizing phenomena at the periphery of the language.

From our experience at SRI with the very broad-coverage grammar DIALOGIC, we believe that it is possible today to build a grammar that has 95% coverage, with some parse, of arbitrary English prose of the sort found in newspaper articles. That is, the desired parse may not be the most highly ranked, but it would be somewhere in the list of parses. We estimate that with parse preference heuristics that have been developed at a number of sites, the parser could rank the desired parse most highly 60% to 65% of the time. It is likely that with the use of probabilistic models and the proper training, this number could be pushed up to 75% to 85%. But the training would require the existence of the broad-coverage grammar.

2. WHAT THE NATIONAL RESOURCE GRAMMAR WOULD BE

The National Resource Grammar should include everything we know how to do well. In particular, it should include the following features:

- Complete English inflectional morphology.
- A very broad grammatical coverage, including all the subcategorization patterns, sentential complements, complex adverbials, relative clauses, complex determiners and quantifiers, conjunction and comparative constructions, and the most common sentence fragments.
- Mechanisms for defining and applying selectional constraints, although the actual ontology would not be provided, since that is too domain-dependent.
- A "quasi-logical form" defined for every construction in the grammar. The quasi-logical form would encode all operator-operand relationships, but not attempt to decide among the various quantifier scope readings. It would be easily convertible into other semantic representations.
- The most commonly used parse preference heuristics.
- An optional routine for pronoun reference resolution according to syntactic or centering criteria.
- An optional routine for quantifier-scope generation, either generating all quantifier scopings from the quasi-logical form, or using various common heuristics for ranking the alternate scopings.
- A lexicon of several thousand words, including examples of all lexical categories and subcategories defined by the grammar.

The grammar should be

- Implemented in a unification grammar formalism.

- As modular as possible, for easy modification.
- As reflective as possible of current linguistic theory.
- As neutral as possible on controversial issues.
- Compatible with the classification scheme used in the Penn Tree Bank.

(The third and fourth of these items exert pressure in different directions, of course, and where the conflict is unresolvable, the fourth should take priority.) The system should include

- An efficient parser, programmed in C for portability.
- Convenient grammar development tools, for users to extend the grammar as required in specialized domains.
- Complete documentation on the grammar and on the algorithms used.

During the development of the National Resource Grammar, it should be continually tested on a large set of key examples. Periodically, it should be tested on sentences taken at random from the Penn Tree Bank. Computational linguists and potential users should be consulted regularly to make certain that the system produces analyses that are maximally useful to others.

3. USES

Among the uses of the National Resource Grammar would be the following:

- To provide a convenient syntactic analysis component for researchers wishing to investigate other problems, such as semantics, pragmatics, or discourse.
- To provide a quick and effective syntactic analysis component for government agencies and members of the LDC and others implementing natural language processing applications.
- To serve as a basis for experimentation with stochastic models of syntactic analysis.
- To serve as an aid in the the annotation of sentences in the Penn Tree Bank and other corpora.
- To serve as a challenge to linguists and computational linguists to handle the various phenomena in better ways.

We believe, on the other hand, that a National Resource Grammar should not in any way be required or imposed on research projects. It should be just what it says—a resource. We believe it should promote rather than retard research on grammar and grammar formalisms.

4. ORGANIZATION OF THE PROJECT

By basing the effort on an existing, very broad-coverage grammar, the development of very nearly the entire National Resource Grammar and its supporting system could be completed in one year. Our guess is that roughly 90% of the phrase structure rules and 70% of the constraints on the rules could be completed in the first year. During the second year, the grammar could be put into the hands of a variety of users, who would be consulted frequently, ensuring that the final product was responsive to their needs.

More specifically, we feel the first year's task could be broken down into six different areas, each representing roughly two months' effort for the implementation of an initial solution. Further development of all aspects of the grammar, especially in response to comments from potential users and an advisory committee of linguists and computational linguists, would continue throughout the two years. Completing the initial implementation in the first year would give the developers sufficient time to respond to this feedback.

The six areas are as follows:

1. A core, skeletal grammar, which would allow the developers to trace out the broad outlines of the grammar and give them a tool for testing further developments.
2. The structure of the noun phrase and adjective phrase to the left of the head, including complex determiner and quantifier structures, and adjective specifiers.
3. The auxiliary complex, noun complements and predicate complements, including cleft and pseudo-cleft constructions.
4. The structure of the verb phrase, subcategorization and sentential complements for verbs and adjectives.
5. Relative clauses and other "wh" constructions.
6. Adverbials and other sentence adjuncts.

Conjunction and comparative constructions would be handled not as a separate item, but throughout the effort. It would be a bad idea, for example, to develop a

treatment of nonconjoined relative clauses in Month 3 and a treatment of conjoined relative clauses in Month 10, because the latter may force a complete rethinking of how the former was done. Similarly, semantic interpretation, the lexicon, mechanisms for selectional constraints, and parse preference heuristics would be implemented and documented in tandem with grammar development.

Each of these phenomena is of course a huge problem, and worthy of years of investigation. However, since at least one treatment of each of the phenomena has already been implemented, and encoding the current best existing treatment is what is required, we are confident such a schedule could be met. However, the developers would have to be very sensitive to black holes, since syntax abounds with them, and more grammar development projects have been derailed by them than have avoided them.

Of course, an effort of this scope could not be done by committee, but it would be extremely useful to have an advisory committee consisting of linguists and computational linguists of a wide variety of theoretical orientations. The advisory committee would be solicited, before each two-month period, for key examples and key treatments of the phenomena. As the initial implementation in each area of the grammar is completed, the results, that is, the rules together with complete documentation, would be circulated to the advisory committee for a critique. Where this critique yielded clearly superior solutions to the problems, those solutions would be incorporated into the implementation.

5. CONCLUSION

There will always be researchers who continue to build their own grammars, as they attempt to work out theories of more syntactic phenomena and to make existing formulations more elegant. But there are a large number of other researchers who are building grammars when they want to be and should be working on some of the less understood problems in natural language processing, or when they have an application that needs to be implemented. As a result, research is retarded and applications are delayed. The availability of a National Resource Grammar would free researchers to push on the frontiers of the field and to move applications into the workplace, rather than duplicating what has been done often before.

After over thirty years of extensive research in linguistics and computational linguistics on the syntax of English, it is time for the development of the National Resource Grammar, reflective of the best that we know and available for general use.