Natural Language Parsing, ed. Dowty, Karttunen, and Zwicky. Cambridge Univ. Press, 1985.

Introduction

LAURI KARTTUNEN and ARNOLD M. ZWICKY

1. Parsing in traditional grammar

Like so many aspects of modern intellectual frameworks, the idea of parsing has its roots in the Classical tradition; (*grammatical*) *analysis* is the Greek-derived term, *parsing* (from *pars orationis* 'part of speech') the Latin-derived one. In this tradition, which extends through medieval to modern times,

(1) parsing is an operation that human beings perform,

(2) on bits of natural language (usually sentences, and usually in written form),

(3) resulting in a description of those bits, this description being itself a linguistic discourse (composed of sentences in some natural language, its ordinary vocabulary augmented by technical terms);

(4) moreover, the ability to perform this operation is a skill,

(5) acquired through specific training or explicit practice, and not possessed by everyone in a society or to equal degrees by those who do possess it,

(6) and this skill is used with conscious awareness that it is being used.

Parsing, in the traditional sense, is what happens when a student takes the words of a Latin sentence one by one, assigns each to a part of speech, specifies its grammatical categories, and lists the grammatical relations between words (identifying subject and various types of object for a verb, specifying the word with which some other word agrees, and so on). Parsing has a very practical function:

> It is not generally realized, even in the schools, how difficult it is for anyone to control the expression and interpretation of language, and that control is as difficult to teach as it is to achieve. The traditional means of teaching control, to pupils at all levels, in their own language as well as in foreign languages, is the set of analytical procedures called grammar.
>
> (Michael 1970:1)

In other words,

(7) the reason for a discipline of parsing is to increase one's mastery over expression in language.

Another important part of the tradition is a separation between grammar and logic. Parsing is analysis for the purposes of grammar; quite a different sort of analysis is appropriate in the study of argument. Although the distinction between grammatical form and logical form has been drawn in a number of ways, not always clearly, it plays a role in linguistic discussions from Aristotle through Port Royal to Chomsky. Here we must stress the fact that, in its traditional sense, parsing is in no way an extraction of properties and relations that are of direct *semantic* relevance. In rather modern phrasing,

(8) the descriptions in (3) are grammatical in nature; that is to say, they describe facts relevant to the co-occurrence of and alternation between units in a particular language.

Note that (8) does not specify any particular theory of grammar; one can parse sentences with respect to any given theory. Indeed, much of the history of parsing until a few decades ago can be understood as the direct consequence of the history of (partial) theories of grammar. Changes in the list of parts of speech, in the list of grammatical categories, or in the list of grammatical relations carry with them changes in what has to be said in parsing a sentence.

We now summarize these eight characteristics of parsing in the Western grammatical tradition. Characteristic (1) says that parsing is done by human beings, rather than by physical machines or abstract machines. Characteristic (2) specifies that what is parsed is a bit of natural language, rather than a bit of some languagelike symbolic system. Characteristic (3) specifies that the analysis itself is a bit of natural language, rather than a bit of some languagelike system, and characteristic (8) that the analysis concerns grammatical rather than logical properties. Characteristic (4) tells us that parsing is heuristic rather than algorithmic, characteristic (5) that it is learned by certain people and not "given" within a society. According to characteristic (6), parsing is overt rather than covert. Characteristic (7), finally, says that the function of parsing is pedagogical.

2. New notions of parsing

In this century the word *parsing* has come to be extended to a large collection of operations that are analogous in some ways to the traditional one just described, but differ from it in one – or usually more – of the eight characteristics. These changes result from a series of new conceptualizations, partially independent of and partially interconnected with one another, in theoretical linguistics, formal language theory, computer science, artificial intelligence, and psycholinguistics. Although the historical roots of these ideas are in some cases fairly deep, they flower together only about the middle of this century, in the 1950s and early 1960s.

3. Parsing in formal linguistics

In linguistics the first of these changes was to view the rationale for parsing not as pedagogical, but rather as scientific – in other words, to emphasize the descriptive, rather than the prescriptive, side of characteristic (7). This shift in emphasis was largely the work of structuralist linguistics, and in its train came a significant change in characteristic (3), as a concern for precision in grammatical descriptions led increasingly to their formalization. The end of this movement away from the informal and discursive descriptions of traditional grammar was a view of these descriptions as completely formal objects – in particular, as *constituent structures* (assigning words to parts of speech and describing which adjacent constituents can combine with one another, in what order they combine, and what phrase category the combination belongs to).

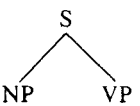
This particular formalization of grammatical descriptions is bought at some cost, for the information coded in constituent structures is considerably less than that supplied in traditional parsing. For example, in such structures heads and modifiers are not systematically marked, discontinuous constituents are not recognized, the relationship between the determined and determining constituents in government and agreement is not generally indicated, and only some of the applicable relations between NPs and Vs are noted. It is striking in this regard to compare the "coverage" of an elaborated traditional approach to parsing, such as Reed and Kellogg diagrams (see Gleason, 1965:142–51 for a succinct presentation), with that of formalized constituent structures (for instance, Harris, 1946 and Chomsky, 1956). Succeeding developments in grammatical theory can, in fact, be seen as attempts to devise fully formalized types of grammatical descriptions with something approaching the coverage of traditional grammars.

Before turning to these developments, however, we must comment on a further conceptual change set off by the move to formalized grammatical descriptions (in particular, constituent structures) as the output of the parsing operation. It is now possible to view parsing as algorithmic rather than heuristic. That is, it is now possible to see the parsing operation as the application of a series of language-particular principles, *phrase structure rules* like NP + VP = S and V + NP (+NP) = VP, to (a representation of) a sentence, in such a way that all the appropriate grammatical descriptions for that sentence, and no others, will be obtained.

Once such a change in characteristic (4) of parsing has been made, the way is open to view principles like NP + VP = S either analytically, as instructions for assigning structures to given sentences in a language, or synthetically, as instructions for composing the sentences of a language. That is, the full set of such principles constitutes a *formal grammar* for the language, which can be seen, indifferently, as having an analytic or *parsing function*, or a synthetic or *generative function*. (Both interpre-

tations appeared early in the history of formal grammatical theory – the generative interpretation most prominently in Chomsky's early work, the parsing interpretation in Hockett's "grammar for the hearer" [1961] and in "dependency grammar" [Hays, 1964; Robinson, 1970; among others].) Indeed, phrase structure rules can also be viewed neutrally, as having a *checking function*, an idea first mentioned in the linguistic literature by McCawley (1968) and developed recently in such work on generalized phrase structure grammar as Gazdar, 1982.

On its analytic or parsing interpretation, the phrase structure rule NP + VP = S licenses the grouping of a constituent known to be (or suspected of being) an NP along with an immediately following constituent known to be (or suspected of being) a VP, into a single constituent of type S; an acceptable constituent structure is then one that is headed by S and can be constructed from entirely by a sequence of such groupings. On its synthetic or generative interpretation, the rule is a *rewrite rule*, customarily formalized as S → NP VP, licensing the replacement of the symbol S in a line of a derivation by the string NP VP, or, equivalently, licensing the branching of a node labeled S in a constituent structure tree into two ordered nodes, labeled NP and VP, respectively; an acceptable constituent structure is then one that can be constructed from the symbol S by such rewriting, or from a node labeled S by such branching. On its neutral or checking interpretation, the rule is a *node admissibility condition*, stipulating that the subtree



is well formed; an acceptable constituent structure is then one that is headed by S and contains only admissible branchings.

When it is recognized that there is more than one way to view the function of phrase structure rules, then it is no longer necessary (though it is not barred) to think of parsing, or for that matter generation or checking, as something human beings do. Instead, these operations can be viewed abstractly, as performed by an idealized device – a change in characteristic (1) of parsing, one that makes characteristics (5) and (6) simply irrelevant.

The consequence of all these reconceptualizations is a distinct second notion of parsing, associated with formal theories of grammar. In this notion

(9) parsing is an operation performed by an abstract device,

(10) on (representations of) sentences in a natural language,

(11) resulting in a formal representation of sentence structure;

(12) this operation is algorithmic.

The next development is for the parsing, generative, and checking functions of a formal grammar for a natural language to be separated. It is a consequence of the particularly simple form of principles like NP + VP = S that they can be interpreted as steps in parsing, as rules for generation, or as node admissibility conditions. But if the steps, the rules, or the conditions are not of this simple, technically *context-free*, form, there is no guarantee that parsing operations, generative operations, and checking operations can be matched in a one-to-one-to-one fashion, and we must contemplate the possibility that the *parser*, the *generator* (sometimes referred to simply as the *grammar*), and the checking device or *acceptor* are three separate devices.

Historically, just such a separation, of parser and generator, followed on the development of transformational grammar as a particular generative framework. And more recently the development of generalized phrase structure grammar has required that generator/parser and acceptor be distinguished. In the first case, the perceived limitations of context-free generative grammar motivated a syntactic theory with at least two distinct components. What is relevant here is that in the new theory constituent structures are not generated directly by rewrite rules, so that a parser cannot be merely a generator run backward. In the second case, the intention was to rehabilitate context-free grammar as a plausible theory of syntactic structure. Part of this program depends on the fact that an acceptor making reference to local context accepts a set of constituent structures generable by a context-free generator (or parsable by a context-free parser); context-sensitive acceptors are thus not simply context-sensitive generators or parsers viewed in a different light.

In general, then, changes in the shape of syntactic theory carry with them consequences, often profound, with respect to the role and nature of a parser. In *monostratal theories* there is only one sort of syntactic representation, and it is the parser's business to assign just the right representations to any given sentence. The representations themselves will of course vary from one theoretical framework to another; the representations of arc pair grammar (Johnson and Postal, 1980) are graphs of a type quite different from the tree structures of classical constituent structure grammar, while the graphs of generalized phrase structure grammar are trees, but trees with node labels decomposed into sets of features (including a slash feature indicating a missing constituent). In *polystratal theories*, with two or more distinct levels of syntactic representation posited, the parser must either construct one level of representation (a *surface*, or *final, structure*) and then translate that into another (a *deep, basic*, or *initial structure*), or it must reconstruct the appropriate initial structures directly, thus operating in a fashion that bears no visible relationship to the operation of the generator.

4. Parsing in formal language theory

Abstracting away from the numerous, undeniably significant, differences in theories of formal grammar and in the parsers associated with them, we observe that the branch of discrete mathematics called *formal language theory*, especially as developed by theoretical computer scientists, has provided most of the conceptual apparatus now current in discussions of parsing, including such fundamental distinctions as those between *top-down* analysis (which begins by examining rules for the top level of structure and attempts to work down to the string of words) and *bottom-up* analysis (which begins by attempting to combine words into larger units and then works "up" to still larger units), between *deterministic* analysis (in which there is only one next step for any particular configuration) and *nondeterministic* analysis (in which a configuration can lead to more than one subsequent step), between *parallel* analysis of alternatives and *sequential* analysis with *backtracking* to alternatives not pursued at earlier stages in analysis, and so on.

Parsing in formal language theory has the same characteristics as parsing in formal linguistics, with the exception of characteristic (10). In formal language theory, the objects of analysis are not (representations of) sentences in a natural language but are instead strings in a symbolic system.

Context-free languages (those with context-free generative grammars) have gotten special attention in formal language theory, and a considerable body of mathematical results on the parsing of them – including upper bounds on the number of steps and amount of storage space for symbols – has been accumulated. In part because there is this body of knowledge in formal language theory, and in part because linguists strive for syntactic theories that are as restricted as possible, the literature on parsing in formal linguistics has been heavily preoccupied with parsing context-free languages.

5. Parsing computer languages

The notion of parsing in formal linguistics and formal language theory closely resembles notions developed in several other contexts. Consider, as a first example, the task confronting the designer of a computer "language." If an actual machine is to do anything with the strings of symbols that serve as its input, it must group them into particular types of "words" and "phrases" that can then serve as signals for specific alterations in the machine's internal states. This is a task quite analogous to the grouping of phonemes or letters into words and phrases, and the assignment of these units to categories, in a linguistic analysis – except that the designer of a computer language is free to *stipulate* the principles of grouping and interpretation (in particular, the designer is free to legislate that all "sen-

tences" and "discourses" are unambiguous in structure and interpretation), whereas the linguist is obliged to *discover* the principles that happen to hold in the (natural) language in question. The closer the designer would like the computer language to approximate the appearance of some existing natural language, the more the designer is engaged in something like linguistic analysis, but even if a particularly simple sort of context-free grammar is stipulated for a computer language, the machine has to do something akin to parsing. That is, for the designer,

(13) parsing is an operation performed by a computer (including both software and hardware),

(14) on symbolic input in a constructed (more or less) languagelike system,

(15) resulting in successive (partial) groupings of symbols into larger units of particular types, and the interpretation of these groups as changes in machine states;

(16) this operation is algorithmic.

Note that in (15) we have left open the possibility that changes in machine states might begin to be effected before a complete parse of the input is constructed. In parsing for computational purposes, as opposed to parsing for the purposes of formal language theory, there are two rather different types of operations: parsing proper, the task of grouping and labeling; and the transformation of symbol groups into real-world effects. Nothing says that the first sort of operation must be completed before the second engages. The two processes might be interleaved.

6. Natural language parsing in artificial intelligence

A second example of an operation analogous to parsing in formal linguistics and formal language theory comes from artificial intelligence (AI), in particular from studies of "natural language understanding" by computer. The machine's task in this case is essentially the same as in the previous section – except that the input is a symbol string in a natural language rather than a (designed) computer language, and also that the so-called understanding might proceed in a strategic rather than algorithmic fashion (hence, might fail for some inputs). Here,

(17) parsing is an operation performed by a computer (including both software and hardware),

(18) on (representations of) sentences in a natural language,

(19) resulting in successive (partial) groupings of symbols into larger units of particular types, and the interpretation of these groups as changes in machine states;

(20) this operation may be either algorithmic or heuristic.

The third characteristic here, in (15) and (19), appears to be the same for parsing computer languages as for parsing in a natural language understanding system, yet it is obvious from reviews of AI approaches (Barr and Feigenbaum, 1981:256–72; Winograd, 1983:chs. 3, 7) that these latter are typically much more complex than the parsing schemes that have been advanced for computer languages. Computer languages have usually been designed as especially simple (and unambiguous) context-free languages, so as to permit rapid interpretation; the historical exemplar is Backus Naur Form in the syntactic definition of ALGOL 60 (Naur, 1963). Natural language understanding programs, even those restricted to small conceptual worlds and reduced means of expression, have to cope with some of the messiness of ordinary language.

In AI parsing, as in parsing computer languages, many options exist for apportioning work between parsing proper and the interpretation of symbol groups, and also for coordinating these two types of operations. Existing AI parsing systems range from those in which parsing proper gives a complete structure of the sort provided in formal grammatical analysis, to those in which parsing proper is reduced to a minimum – in the extreme case, to the recognition of a few key words. AI parsing schemes can also be devised on principles quite different from those of formal grammars, for instance on the basis of pattern-matching routines. And as before, interpretation might follow parsing proper, or (as is very common these days) take place in tandem with in and interacting with it.

Note that AI parsing shares with computer-language parsing, but not with formal-language parsing, the central role given to interpretation. Formal language theory assumes a clear separation between syntax on the one hand and semantics and pragmatics on the other, and it concerns itself only with the former. In computer language design a clear separation is usually assumed, but interpretation is the dominant concern. In approaches to natural language understanding by computer, it is not axiomatic that a clear separation be made, and interpretation is unquestionably the dominant concern.

Note finally that approaches to parsing based on formal language theory are extraordinarily restricted in scope.

a. They do not embrace an explicit semantics.
b. They do not specify, even informally, the relationship between syntactic rules/representations and principles of semantic interpretation.
c. The sentences analyzed exist outside any social context, having no identifiable source or audience, so that there is no body of mutually assumed knowledge with respect to which interpretation can proceed, nor is there any basis for calculating the intentions of a source or effects upon an audience.
d. The sentences analyzed are not anchored in any spatial or temporal context, so that deictic elements cannot be interpreted.

e. The sentence is the largest unit of analysis, so that formal or semantic organization in larger discourse units is not recognized.
f. The word is the smallest unit of analysis, so that meaningful structuring below the word level is not recognized.
g. It is assumed that the chunks listed in a lexicon are word-sized, so that no allowance is made for multiword idioms and formulas.

Undoubtedly this list could be extended, but it will do for now. Computer-language parsing must remedy at least the first two defects, and AI parsing schemes must remedy all seven (or explicitly recognize that they are significantly idealizing away from natural language).

7. Parsing in psycholinguistics

Psycholinguistics provides a third situation in which an operation analogous to parsing in formal theories of grammar might arise. Psycholinguistics proposes to supply an account of mental processes involving language, in particular the processes at work in the perception and comprehension of language, the production of language, and memory for language. It is natural to think of perception and comprehension as including analogues of the parsing operations of formal grammars, and so to view AI parsing schemes as potential models of (portions of) some mental processes. In this view (compare characteristics (1)–(7) listed at the beginning of our discussion of parsing),

(21) parsing is an operation that human beings perform,

(22) on bits of natural language (sentences and discourses, sometimes in spoken form, sometimes in written form),

(23) resulting in mental representations of (aspects of) the syntactic structure associated with these bits, and their interpretation as alterations in other mental representations;

(24) this operation is heuristic (on occasion, it can fail to assign structures or interpretations, or assign the wrong ones),

(25) acquired without specific training or explicit practice, and possessed in some form by everyone in a society,

(26) and used tacitly, rather than with conscious awareness.

As in AI parsing, the defects a–g of formal-language approaches to parsing must be remedied in any extended model of perception and comprehension. Also as in AI parsing, there is a serious question as to how syntax-driven the processes of interpretation are. Does interpretation temporally follow parsing proper, or do they operate interactively and in tandem? Is parsing even necessary at all? In AI work the answers to these questions are in a sense a matter of taste; an AI system has only(!) to mimic aspects of human behavior, not (necessarily) to mirror actual men-

tal organization and functioning, so that a wide variety of approaches are a priori available. In psycholinguistics these are empirical questions, with answers that are, in principle, discoverable. Consequently, psycholinguistic research on parsing has been preoccupied with issues of mental representation: For which hypothesized units is there evidence that they are accessed or constructed during comprehension? For which hypothesized processes is there evidence that they take place (hence, take time) in comprehension?

While there are no universally accepted answers to this last set of questions, it is clear that a model in which parsing proper is completed before interpretation is undertaken has little to recommend it. At the very least, we must assume that as the (surface) constituents of a sentence are identified they are, at least partially, interpreted. There is abundant evidence both that some (surface) constituents are identified during comprehension and that some interpretive processes take place before a sentence is completed (see the survey in Clark and Clark, 1977:45–57). Beyond this, psycholinguists disagree as to whether comprehension is primarily syntax-driven, with the dominant strategies those of parsing proper, or primarily meaning-driven, with the dominant strategies those of interpretation (semantics and pragmatics taken together); and of course they differ on details (for a survey, again see Clark and Clark, 1977:57–79).

The phenomenon of *ambiguity* has come to play a special role in these discussions, because it appears to offer a testing ground on which to compare alternative accounts of language comprehension. If account A predicts that two distinct (partial) representations are constructed during the comprehension of an ambiguous sentence, while account B predicts that only one corresponding representation is constructed, then it ought to be possible to detect the extra time required for the construction of representations in account A as against account B. It ought also to be possible to detect (traces of) the presence of the extra material accessed during the construction of representations in account A as against account B. Ambiguity might then provide a small window into the workings of the parsers we all use.

8. Summaries of the papers