A Descriptive Approach to Language-Theoretic Complexity

James Rogers
(Earlham College)

Stanford: CSLI Publications and FoLLI
(Studies in logic, language and
information, edited by Robin Cooper
and Maarten de Rijke), 1998, x+205 pp;
distributed by Cambridge University
Press; hardbound, ISBN 1-57586-137-2,
$59.95; paperbound, ISBN
1-57586-136-4, $22.95

Reviewed by
Philip Miller
Université de Lille III
and Geoffrey K. Pullum
University of California, Santa Cruz

This interesting and original work applies descriptive complexity theory to the study of natural languages. The first part of the book (64 pages) characterizes the strong generative capacity of context-free grammars in logical terms. The second part (122 pages) discusses the generative capacity of government and binding (GB) theories. The book deserves the close attention of computational linguists, and also of GB theorists who wish to gain a better understanding of the formal properties and consequences of their theory.

Mathematical linguistics was founded on the enormously influential results published by Chomsky in 1959. (We save space by not repeating bibliographical items: our reference list includes only items that Rogers does not cite.) Thus it was based on the theory of rewriting systems and automata. But theoretical computer science and linguistics have diverged since 1959, and the emergence of descriptive complexity theory is one of the results. This field is most often taken (as, for example, by Immerman [1998]) to originate with Fagin's beautiful result (1973, 1974): the class of graph-theoretic problems solvable in polynomial time on a nondeterministic Turing machine is exactly the class that can be stated using a second-order existential sentence—that is, a sentence of the form $\exists X_1 \ldots \exists X_n[\phi]$, where the X_i are variables over properties or relations and ϕ contains only first-order quantifiers.

Fagin's work established a new way to measure computational complexity: instead of asking for a measure of how many tape squares or Turing machine operations are needed for the computation that solves some problem, one can ask instead: "How rich a logic it would take to state this kind of problem?" This has led to a significant new branch of theoretical computer science. It has antecedents in earlier research. In the late 1950s it was proved by J. R. Büchi and by C. C. Elgot that the sets of strings definable by existential sentences of monadic second-order logic (MSO) were exactly the regular languages; that is, a set of strings is regular if and only if it is a model of a sentence of the form $\exists X_1 \ldots \exists X_n[\phi]$, where the X_i are variables over properties of elements, e.g., the property of being an a or being a b, and there is a strict ordering on the elements. Between 1967 and 1970, work by James W. Thatcher, J. B. Wright, Michael O. Rabin, and John Doner extended this result in various ways, considering trees as models for MSO sentences, and from this work it emerged that the sets of finite labeled trees definable in MSO are exactly those that can be generated by context-free grammars. (Engelfriet [1991] presents an elegant further generalization of the models for MSO, this time to arbitrary graphs; the results on trees follow as a special case.)

Rogers connects this work to the post-1980 trend in theoretical linguistics focused on grammars composed of (or incorporating) declarative statements about the syntactic structure of natural language expressions. Linguists of most theoretical persuasions today talk about languages in a way that is at least partially reminiscent of the work of the computer scientists mentioned above: they state (however informally) conditions applying directly to syntactic structures. For example, consider a statement of binding theory in GB, such as "An anaphor is governed in its binding domain." This is a condition on sentence structure, not a component of a device for building structure. A representation that displays binding relations either satisfies it or does not. Description in such terms can be called *model-theoretic syntax* (the term appears to originate with Rogers's title for a class he gave at the European Summer School in Logic, Language, and Information in 1996). By contrast, the dynamic structure-building syntax of Post production systems and transformational grammars can be referred to as *rewriting* syntactic description.

But most current theories are actually hybrids of model-theoretic and rewriting syntax. For example, the "Move" operation of GB and the more recent minimalist program is interpreted as a schema over instructions for building objects: "Move" is not a statement that is true or false of any given structure. Rogers is interested in the program of describing syntactic structure in a way that is purely and rigorously model-theoretic (as hardly any work in linguistics has been; Rogers [1997] correctly notes Johnson and Postal [1980] as one of the few exceptions).

Rogers defines a special-purpose MSO language, $L^2_{K,P}$. The intended models are trees. He proves that ϕ is a sentence of $L^2_{K,P}$ if and only if the set of all the finite models of ϕ is the set of trees generated by some context-free grammar. (The "only if" direction of the proof actually requires a more complex condition involving the projection of a set of trees generated by a finite set of CFGs; see pages 56ff.) Chapter 4 is devoted to a careful comparison of $L^2_{K,P}$ with *SnS*, a logic studied earlier by Rabin, and Rogers shows that in a specific sense they are strongly equivalent. Thus there is an exact descriptive-complexity characterization of the context-free languages. And it is a characterization in terms of strong rather than weak generative capacity.

What makes $L^2_{K,P}$ a particularly desirable description language for natural languages is that (as in effect shown earlier by Rabin) its satisfiability problem is decidable: it can be decided algorithmically whether there is a model that satisfies a given theory. Notice that we do not have the analogous property for (e.g.) context-sensitive grammars or transformational grammars: there is no algorithm that can take a grammar of either of these types and determine whether there might be some sentence that it generates, because the emptiness problem for both is (by an easy application of Rice's theorem) undecidable.

The second part of Rogers's book applies $L^2_{K,P}$ to the analysis of GB, roughly as assumed in Chomsky's 1986 book *Barriers*. The investigation yields several interesting definability results. Rogers shows that all of the most central parts of GB theory—X-bar theory, the lexicon, theta theory, Case theory, binding, control, leftward and rightward phrase movement, chains, the empty-category principle, head movement, reconstruction—can be defined in $L^2_{K,P}$. Thus "the language licensed by a particular theory within the GB framework is strongly context-free" (p. 6). This was not appreciated during the 1980s, when GB was in competition with generalized phrase structure grammar (GPSG), and GB proponents insisted that GPSG's claim of context-freeness would prevent it from capturing the facts of natural languages whereas GB did not have this drawback. (A similar development occurred when it was proved by Nozohoor-Farshi [1987] that the Marcus [1980] parser, touted as an important existence proof for parsing of transformational grammars, in fact only had weakly context-free recognizing power.)

Importantly, these results do not show, like so many such results, an excess of expressive power in a formalism previously thought to be restrictive. Here a theoretical system turns out under analysis to be more restrictive than it was thought to be. Although GB (in the version Rogers considers) employs transformations, the "structure-preserving" use that is made of them in the description of the core structures of English makes it possible to represent a sentence as a single annotated s-structure tree that contains the corresponding d-structure as a subtree. Rogers shows this by recasting movement in terms of relations between traces (some of which have phonologically unrealized daughter material referred to as "phantom constituents").

Dutch cross-serial dependencies and Swedish unbounded dependencies cannot be described in $L^2_{K,P}$. Rogers diagnoses the reason as the same in both cases: description of them would require the existence of unboundedly many "chains" crossing a given node. But interestingly, Manzini and Stabler have independently argued for a constant bound on overlapping chains (Manzini's locality theory implies the maximum number of chains overlapping at any given node is 2), which means that they have (apparently unwittingly) imposed a theoretical restriction on GB that would limit it to the context-free languages, making languages like Dutch, Swiss German, and Swedish impossible to describe.

There are some real differences between the expressive power of GB theories and context-free grammars, however. Two theoretical devices assumed in recent GB go beyond strongly context-free power. Movement by copying, advocated in various works (notably Chomsky's minimalist program) is not definable in $L^2_{K,P}$. And the most interesting nondefinability result Rogers offers concerns free indexation, assumed widely in GB since about 1980. Free indexation assumes that phrases (specifically noun phrases) are freely assigned numerical indices. In other words, a node is taken to be labeled by both a category and a random integer called the index. Syntactic and semantic principles are permitted to refer to coindexing. Rogers defines an extension of $L^2_{K,P}$ that incorporates free indexing and a dyadic predicate that holds of two nodes, n labeled by k and n' labeled by k', if and only if $k = k'$. He then shows (by reduction to the origin-constrained tiling problem) that the resultant theory is undecidable. Thus the device of coindexing (not present in GPSG) increases expressive power to the extent that there is no general algorithm to determine whether a given theory is satisfiable.

$L^2_{K,P}$ gives a characterization of the strong generative capacity (SGC) of context-free languages, in the sense of Miller (1999), since it provides a logical characterization of the properties that are assigned to a string in virtue of the fact that it has a given derivation. Thus it does in a sense show a certain GB theory to be strongly context-free (p. 6). But that statement must be understood as based on a more classical definition of SGC, involving uninterpreted sets of structures: one can provide sets of annotated trees that are the ones GB grammars would generate, and there is a proof that those sets can be generated by context-free grammars. No logical interpretation of those properties of GB theory coded in the indices (for example) is provided, or can be provided, as Rogers notes. He remarks on page 70 that "if descriptive complexity results similar to ours can be established for larger language-theoretic complexity classes (for the indexed languages, for instance) it may be possible to make the argument that GB can be restricted to principles that are definable in the corresponding theory without losing the ability to account for the full class of natural languages." In other words, what has been done for the (strongly) context-free languages might be also possible for larger classes, like the TAG languages or the indexed languages (Rogers [1998] shows that in fact this can be achieved for TAGs). We believe it would be a fruitful line of research to try to further extend Rogers's logical analysis so that it can provide explicit interpretations for linguistic properties that go beyond domination, linear order, and constituency. This would amount essentially to integrating some of the ideas of Miller (1999) into Rogers's more rigorous and well-defined logical framework.

Still, Rogers's results as they stand are important and exciting. Techniques for showing exactly where the excessive descriptive power of a theory might be located are clearly of great value to anyone, of any theoretical leaning, who wishes to understand syntactic theory deeply. For instance, when one considers the analysis of X-bar theory proposed in Kornai and Pullum (1990), it becomes clear that among the six defining properties of X-bar theory, five are easy to express in $L^2_{K,P}$, whereas the sixth, optionality (which claims that the nonhead daughters in any subtree are optional), is literally impossible. Rogers's approach makes it clear why: Rogers assumes a domain of nodes and various properties and relations. In a model of an $L^2_{K,P}$ sentence, the properties and relations are assigned in a way that determines treehood. But defining optionality requires quantifying over sets of trees. The English phrase "the daughter node α in subtree τ is optional" might appear to say something about τ, but when things are formalized in Rogers's terms we can see that it does not: the assertion it makes is about a set of subtrees that τ belongs to. We can neither verify it nor refute it by examining τ.

It is of interest, therefore, that Kornai and Pullum (1990) point out that optionality is a condition that has only ever been paid lip service by linguists. In practice no syntactician has even seriously attempted to maintain it. It is incompatible with the existence of strict transitive verbs, or with the existence of languages such as English in which both the predicate (verb phrase) and the subject (noun phrase) are obligatory in the finite clause. The inability of $L^2_{K,P}$ to state this principle turns out to be a virtue, if the practice of syntacticians is to be trusted.

Rogers's work has already been influential, clearly inspiring most of the work published in Kolb and Mönnich (1999). The most clearly related work within computational linguistics is probably that of Patrick Blackburn on applying modal logic to the description of trees (the work of Blackburn and his associates on describing trees in logical terms—modal logic, in their case—has been developing since about 1993; see Blackburn and Meyer-Viol [1997] for an overview). It should also be noted that in research not included in the book under review, Rogers (1997) has intensively studied the GPSG framework, showing that significant advantages accrue from reformulating the 1985 GPSG theory of Gazdar, Klein, Pullum, and Sag in terms of direct statements about trees in a logical language for tree description, rather than through the rewriting system-derived techniques standardly employed in GPSG. Again this yields interesting insights into the content of the theory, and the first part of this book would be a very useful guide for anyone consulting the 1997 paper.

In sum, we recommend to linguists interested in achieving a clearer and sharper view of natural language syntax that they should make sure they do not overlook this book.

References

Blackburn, Patrick, and Wilfrid Meyer-Viol. 1997. Modal logic and model-theoretic syntax. In M. de Rijke (editor), *Advances in Intensional Logic*, pages 29–60. Kluwer Academic, Dordrecht, Netherlands.

Engelfriet, Joost. 1991. A regular characterization of graph languages definable in monadic second-order logic. *Theoretical Computer Science*, 88:139–150.

Fagin, Ronald. 1973. *Contributions to the Model Theory of Finite Structures*. Ph.D. dissertation, University of California, Berkeley.

Fagin, Ronald. 1974. Generalized first-order spectra and polynomial-time recognizable sets. In Richard Karp, editor, *Complexity of Computation, SIAM-AMS Proceedings*, 7:27–41.

Immerman, Neil. 1998. *Descriptive Complexity*. Springer Verlag, New York.

Johnson, David E. and Paul M. Postal. 1980. *Arc Pair Grammar*. Princeton University Press, Princeton, NJ.

Kolb, Hans-Peter, and Uwe Mönnich, editors. 1999. *The Mathematics of Syntactic Structure: Trees and Their Logics*. Mouton de Gruyter, Berlin.

Kornai, Andràs, and Geoffrey K. Pullum. 1990. The X-bar theory of phrase structure. *Language*, 66:24–50.

Marcus, Mitchell. 1980. *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge, MA.

Miller, Philip. 1999. *Strong Generative Capacity: The Semantics of Linguistic Formalisms*. CSLI, Stanford, CA.

Nozohoor-Farshi, R. 1987. Context-freeness of the language accepted by Marcus' parser. *Proceedings of the 25th Annual Meeting*, pages 117–122. Association for Computational Linguistics.

Rogers, James. 1997. "Grammarless" phrase structure grammar. *Linguistics and Philosophy*, 20:721–746.

Rogers, James. 1998. A descriptive characterization of tree-adjoining languages (project note). *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Volume II, pages 1,117–1,121.

Philip Miller holds a Ph.D. in linguistics from the University of Utrecht (Netherlands) and has published three books and numerous papers on topics in mathematical, theoretical, and descriptive linguistics. He is a professor of linguistics at the Université de Lille 3 (France) and director of the SILEX research unit of the Centre National de la Recherche Scientifique. Miller's address is SILEX, Université Charles de Gaulle Lille 3, B.P. 149, 59653 Villeneuve d'Ascq, France; e-mail: pmiller@ulb.ac.be.

Geoffrey K. Pullum earned his Ph.D. in general linguistics at the University of London and has published widely in mathematical, theoretical, and descriptive linguistics. He is a professor of linguistics at the University of California, Santa Cruz, and an associate member of the SILEX research unit of the CNRS. Pullum's address is Cowell College, University of California, Santa Cruz, Santa Cruz, CA 95064; e-mail: pullum@ling.ucsc.edu.