Marcus Kracht, *The Mathematics of Language* (Studies in Generative Grammar, 63). Berlin: Mouton de Gruyter, 2003; xvi + 589pp. US $127.40.

Reviewed by Geoffrey K. Pullum

Mathematical tools and concepts find many applications within the linguistic sciences: the fast Fourier transform is crucial tool for acoustic phoneticians working on speech analysis; contemporary computational linguists make heavy use of probability theory for stochastic modeling; experimental psycholinguists and sociolinguists use statistics; and so on. But Marcus Kracht's book *Mathematics of Language* is not concerned with topics of this sort. For the last fifty years a large amount of research within pure linguistics, particularly in the United States, has been based on separating natural languages from the physical, social, and cultural matrix in which their use is embedded, and regarding them in idealized form as collections of structured mathematical objects. This has led to the application of algebraic methods — the discrete mathematics underlying formal language theory, computational complexity theory, and mathematical logic — to the description of natural languages and the study of their abstract properties.

In the simplest case, the objects in question are taken to be just finite strings over some kind of finite vocabulary of words or letters (I shall use the term *stringset* for a set of finite strings.) Under the standard view, the sets involved are taken to be infinite, since nothing in grammar seems to suggest any definite cutoff length. An infinite set of strings S can be studied by trying to devise a formal system that exactly generates S, or by constructing an abstract computing machine that accepts all and only the members of S, or by writing a set of statements in a logic that has S as the set of all of its finite models. All three perspectives find a place in this book.

Kracht formerly held mathematics positions at the Freie Universität Berlin and the Universität Potsdam, and is now at the Department of Linguistics at UCLA. He has published significant work in modal logic, mathematics, and theoretical computer science as well as all the central areas of linguistics (phonology, morphology, syntax, semantics). The book he has written is an advanced introduction, and it is the best and most authoritative book on its subject matter that I have seen, with excellent coverage of issues relevant to contemporary linguistics. It is often highly original in its content and presentation.

There is certainly a need for a new text on the algebraic tools that are relevant to linguistics. W. J. M. Levelt's truly excellent 3-volume 1974 textbook [6] had remarkably wide coverage (Levelt's psycholinguistic interests lead him to cover work on 'learnability', also known as grammar induction, which Kracht does not touch on), but sadly has long been out of print. And the standard text by Partee, ter Meulen and Wall [9] is now more than fifteen years behind the leading edge of research, especially with respect to grammars and automata. (Though it was published in 1990, the Partee el al. volume reports as open the question of whether the complement of a context-sensitive stringset is always context-sensitive, which was settled in the affirmative in 1987, at Partee's institution!) Though strong on formal semantics, it completely misses important topics in other areas (parsing and computational complexity, for example), and it looks positively fusty beside Kracht's much more up-to-date and considerably more mathematical book.

The decidedly mathematical cast of Kracht's book should be stressed. Readers of *The Mathematical Intelligencer* will probably get on with it well enough, but others should be warned that Kracht assumes a lot of mathematical sophistication: graduate students whose first degree is in humanities or social science may experience symbol shock. Kracht does not pamper those who crave intuitive presentations. He will not explain that a finite automaton accepts exactly those strings on which there is a run beginning in the start state and ending in a final state; he will expect you to see that immediately when he tells you (on p. 96) that $L(\mathfrak{A}) = \{\vec{x} : \delta(\{i_0\}, \vec{x}) \cap F \neq \emptyset\}$. (And yes, those unacquainted with old-style German blackletter 'Fraktur' characters $\mathfrak{A}, \mathfrak{B}, \mathfrak{C}$, etc., will need to brush up on them, for Kracht often draws on Fraktur when naming variables that range over higher types.)

The 553 pages of Kracht's main text are organized into six fat chapters. Here is a rough-and-ready guide to the chapter contents:

- Ch 1, "Fundamental Structures" (pp. 1–93), does a rapid review of algebra, strings, levels in linguistics, graphs, trees, the hierarchy of rewriting systems, constituent structure representation, and Turing machines;

- Ch 2, "Context Free Languages" (pp. 95–175), treats regular sets, finite automata, normal forms, recognition, ambiguity, parsing, semilinearity, and the issue of whether natural languages (when idealized as stringsets) are context-free in the formal language theory sense;

- Ch 3, "Categorial Grammar and Formal Semantics" (pp. 177–280), introduces propositional logic, lambda calculus, categorial grammars, the Lambek calculus, and the rudiments of Montague semantics;

- Ch 4, "Semantics" (pp. 281–365), covers intensionality, binding, quantification, cylindric algebras, more Montague semantics, presupposition, and ternary-valued logic;

- Ch 5, "**PTIME** Languages" (pp. 367–459), discusses a variety of topics mostly concerning the family of stringsets that Aravind Joshi [5] refers to as 'mildly context-sensitive'; and

- Ch 6, "The Model Theory of Linguistic Structures" (pp. 461–553) takes a model-theoretic look at the description of strings, phonological units, categories, trees, command relations, and transformational derivations, as well as modern (post-1980) phrase structure grammars and transformational grammar.

Several chapters have long, dense sections where Kracht takes us through full-dress proofs of theorems he considers particularly important:

- in Ch 2, Parikh's theorem (that a stringset over the natural numbers is semilinear iff it is letter-equivalent to a regular set) and the Ginsburg-Spanier theorem (that a stringset over the natural numbers is semilinear iff it is definable in Presburger arithmetic);

- in Ch 3, the rather inaccessible 1997 proof by Pentus that Lambek-calculus grammars define only context-free stringsets;

- in Ch 4, Gödel's completeness theorem for first-order logic;

- in Ch 5, the theorem of Chandra, Kozen and Stockmeyer that **PTIME** (the set of functions computable in polynomial time by a deterministic Turing machine) is exactly **ALOGSPACE** (the set of functions computable by a logarithmically space-bounded alternating multitape Turing machine); and

– in Ch 6, Büchi's theorem (that a stringset is a regular stringset iff it corresponds to a class of stringlike structures that is finitely axiomatizable in weak monadic second-order logic, henceforth wMSO) and Doner's theorem (a direct analog of Büchi's theorem holding for trees rather than strings).

There are relatively few linguists who have the background to read and understand the proofs of these theorems, or even to find them in the mainly mathematical sources where they first appeared. Kracht offers clear and rigorous presentations of the results and the proofs all gathered together in one place.

No previous mathematical linguistics text has covered the theorems of Büchi and Doner, but I think Kracht is quite right about their significance for linguistics. They are part of the foundations on which descriptive complexity theory has been built within computer science (see Immerman's survey [4]), and that perspective is beginning to yield new kinds of linguistic insight ([13], [14]).

Notice also that two chapters out of Kracht's six are devoted to semantics. The book rightly makes an effort not to limit attention to the idealization (and oversimplification) that equates languages with stringsets, in the way computer scientists standardly do; he considers techniques for representing languages as sets of structures more complex than strings (trees or graphs of some other kind), and he also considers techniques for representing languages as systems of signs — meaning-bearing expressions with both a formal and a content-conveying aspect. (He announces in the Introduction (p. ix) that he assumes there is no "fundamental difference between natural and nonnatural languages" — the position adopted by Richard Montague [8].) The material on semantics sheds some further light on one of the most important general characteristics natural languages are claimed to possess, namely compositionality — the property of being designed in such a way that the meaning an expression has is fixed as a function of nothing more or less than the meanings that its parts have.

The book is handsomely bound, and Kracht's LaTeX typesetting work is excellent throughout — fonts and symbols are used with skill, taste, and consistency, and various clever but unobtrusive notational devices not seen in other mathematical linguistics books (like consistently using α for a variable over symbols in an alphabet A but $\vec{\alpha}$ for a variable over strings in A^*) are deployed with care. But as always with an author-typeset book, there are slips. The reader will be occasionally distracted by errors in English grammar. I don't mean purist shibboleths; I mean small but real departures from idiomatic English like using the present perfect with an agent phrase denoting a deceased person ("Emil Post... has shown" on p. 65 and similar uses passim); using *allow* with a subjectless infinitival complement ("allows to derive the claim" on p. 469 and other such uses passim); "so must be *F*" for "so must *F* be" on p. 55; "can obviously not" for 'obviously cannot' on p. 285 and "can otherwise not write" for "cannot otherwise write" on p. 510; and so on.

Let me note also a pet peeve about mathematical style. Kracht's Theorem 5.11 (p. 375) says simply "ξ *is in* ***PTIME***"; but ξ is merely an ad hoc symbol for a function that serves a temporary purpose in simplifying the proof of the Chandra/Kozen/Stockmeyer theorem (it receives its rather tricky definition in the middle of the text on p. 375). After a couple of mentions in the proof, and one in Lemma 5.21 on p. 380 ("ξ *is in* ***ALOGSPACE***"), it has no further role and disappears from the exposition. Thus Theorem 5.11 cannot be quoted out of this book: it is unintelligible outside the context of the specific section in which it is stated. I have occasionally seen mathematicians doing this sort of thing elsewhere, but I want to register my negative opinion of the practice anyway. No nonce variables in the statements of theorems. I don't want to open the paper someday and learn that someone has won the Fields Medal for proving that x has the property P!

One typographical error with impact on intelligibility is worth correcting: on p. 61 we read that "Context sensitive grammars are contracting." After a few seconds of puzzlement the attentive reader will realize that Kracht must mean the exact opposite: context-sensitive grammars in the mathematical sense are non-contracting (the rules have the form 'φ may be replaced by ψ' where ψ is never shorter than φ). Real errors in mathematical content are quite rare, though, as far as I have been able to determine. A few of the errors (only five as this review went to press) are already known to Kracht, who has begun an errata list at `http://www.linguistics.ucla.edu/people/Kracht/html/remarks/mol.html`; I will not reproduce all of the content of that page here, but I will say something about a couple of points that connect with interesting issues.

The most serious mistakes in the book are in the brief section on the "index languages." Kracht means the class of stringsets defined by Aho [1] and named the "indexed languages" — henceforth, ILs. (Kracht does not cite Aho, or any other works on ILs, and his neologism "index languages" is an annoyance for anyone doing a literature search, since nobody else uses this term.)

The ILs are generated by indexed grammars (IGs), which constitute a generalization of context-free grammars that in effect gives them an infinite set of nonterminal symbols. Nonterminals can bear strings of marks called indices, and rules can stack new indices onto a nonterminal or pop indices off. The IGs define a large family of stringsets — it is a proper subset of the context-sensitive and a proper superset of the context-free, and identical with the set of string yields of those tree sets in which the node-to-root path sets are context-free. The ILs have the same sort of mathematically desirable properties that make the CF stringsets such a mathematically robust and interesting family, yet it is straightforward to give IGs for many classic cases of context-sensitive stringsets that are not context-free. The following four stringsets, for example, are all ILs:

$$\{a^{2^n} : n \geq 0\}$$

$$\{a^n b^n c^n d^n e^n : n \geq 0\}$$

$$\{cw^n : w \in (\mathbf{a} \cup \mathbf{b})^*\} \quad \text{for any } n$$

$$\bigcup_{n \geq 0} \{cw^n : w \in (\mathbf{a} \cup \mathbf{b})^*\}$$

However, is is known as a consequence of a pumping theorem [3] that the following are not ILs:

$$\{a^{n!} : n \geq 1\}$$

$$\{(cw)^{|w|} : w \in (\mathbf{a} \cup \mathbf{b})^*\}$$

There are well-known and difficult open problems about whether certain stringsets are ILs; the permutation closure of $\{a^n b^n c^n : n \geq 0\}$ is one example.

The major mistake Kracht makes is discussing the ILs in the chapter on **PTIME** languages. Kracht asserts (pp. 425–426), but does not prove, that the IL recognition problem is easily shown to be in **PTIME**. The claim is not true [16].

Kracht also sets it as an exercise for the student (p. 433) to show that the set of all formulae of first-order predicate logic in which every quantifier binds at least one variable cannot be an IL. It would be a promising student who did this as a homework exercise; the conjecture, publicized by

Bill Marsh and Barbara Partee more than twenty years ago [7], is another open problem about ILs, open for over twenty years and regarded as difficult.

A couple of small mathematical errors occur in §2 of ch 6 (pp. 470–472). Definition 6.2 on p. 470, should have said "L is an arbitrary finite set" rather than "L is a nonempty finite set", and the statement on p. 472 that "models are only defined on nonempty sets" should have been dropped. Moreover, the statement that an empty structure would be "a model of every formula" is false. It would be true under the modal logic that Kracht uses (p. 473) to accomplish part of the subsequent proof (an arbitrary modal logic formula φ holds in an empty structure M since, by vacuous satisfaction of the universal quantifier, M satisfies φ at n for every node n in M); but this doesn't hold in the context of p. 472, where the topic is wMSO.

There is indeed an inelegant little exception to Büchi's theorem for those model theorists who take a structure to have a nonempty domain by definition: a regular expression like $\mathbf{a}^*$, denoting the set of all possible finite strings of *a*s, does not correspond to the set of string models of any wMSO formula, for ε (the unique string of *a*s that contains zero symbols) is in the set but its analog would be the putatively illegal unique empty stringlike structure. It seems to me that the best solution is to allow empty domains in the definition of 'structure'. Empty domains raise no real problem in the context of wMSO. All positive existential statements will be false in the empty structure, and all positive universal statements true, but this is not problematic. Describing a stringset not containing the empty string would entail having at least one non-vacuous existential statement in the description ($\exists x[x = x]$ would do). All positive universally quantified statements would be true in the empty structure, but no real problems arise from this. The parallel between model-theoretic, automata-theoretic, and grammar-theoretic results becomes cleaner and clearer. The only reason Kracht cannot take this course is that he wants to use quantified modal logic in constructing his proof, and for a modal logic there really is a problem with the empty structure.

An additional small inaccuracy is found on p. 530: according to Kracht, the unbounded branching permitted in generalized phrase structure grammar [2] "puts it just slightly outside of context freeness"; but this is not true. Context-freeness is a matter of stringset definability. Unbounded branching allows arbitrarily long coordinate constituents, like the English noun coordination '`red, orange, yellow, green, blue, indigo, violet, teal, and puce`', to be given n-way branching tree structures that do not arbitrarily group the conjuncts into pairs. But every context-free stringset definable with unbounded branching is also definable without it. What Kracht probably means is that assigning phrase structure with arbitrary branching to a CF stringset is not possible for a context-free grammar. It is possible using model-theoretic techniques, though, as shown in a beautiful and original discussion by Rogers [15].

One section in Kracht's book falls way below the standards set by the rest, and it is not entirely his fault. In §5 of ch 6 (pp. 515–529) he provides a brief introduction to some aspects of modern transformational-generative grammar (TG). It is a rather sad fact about current theoretical linguistics that the mathematical rigor to which TG once aspired has been eroding away for some time. What originated in some fairly precise ideas about formal systems (taken from the work of the mathematician Emil Post in the 1940s) is today presented as a jumble of unexplicated metaphors, undefined phraseology, and misleading allusions to earlier work. Writers who try to write material explaining it to new audiences frequently fall into repetition of hackneyed and sometimes incoherent set phrases.

Kracht's §6.5 simply adopts the prevailing tone. Jettisoning his earlier insistence on algebraic exactitude, he takes to repeating familiar phrases of obscure import, without elucidation. Take this randomly selected passage:

> V only stands for the verb, not for the complex consisting of the verb and the auxiliaries. So, we have to change the SD in such a way that it allows the examples above. Further, it must not be disturbed by eventually intervening adverbials. (p. 516)

At the very least, a mathematician might look for clarification of 'stands for', 'verb', 'complex', 'auxiliaries', 'SD', 'allows', 'disturbed', 'eventually', 'intervening', and 'adverbials'. It is not supplied. 'SD' (which stands for 'structural description') is a particularly important technical term, but no explication is given. (It is not that easy to give one. The formalizations that were given of TG in the 1960s were extremely complex. And incidentally, results based on them [10] showed that TG was just a highly ornate and obscure equivalent of Post production systems.)

The obscurity persists: phrases like "moves into second place" and "contribute to the meaning" are never clarified; "binding is computed" apparently does not really refer to computation; and so on. The mathematician reader might imagine that there is some deep and rigorous theory lying behind these phrases, but there is not. Kracht is simply repeating vague phrases from syntax textbooks, which in turn paraphrase things said in the increasingly allusive journal articles on TG.

Attempting to motivate the notion of 'traces' (p. 519), Kracht asserts that "the place from which an element has been moved influences the behaviour of the structure" (structures, hitherto algebraic objects as defined on p. 15, are now tacitly reconceptualized as having "behaviour" that can be influenced by "places" from which things have been "moved"), and in an effort to clarify this he gives an example involving the distribution of American English *wanna*. Briefly, many American speakers agree that in '`What do you want to do?`', the '`want to`' part can be pronounced as an amalgam that novelists spell as '*wanna*', but in '`What do you want to happen?`' the amalgam is not used. The question is how to account for this. Seeing the topic dragged out once again was particularly depressing for the present reviewer, who has shown that the analysis purportedly supporting 'traces' does not work [11], and elsewhere has presented a superior analysis that does work [12].

There are other points in the section where Kracht does not have the facts straight. He says: "`Himself` is a subject-oriented anaphor . . . it refers to the subject of the same sentence." This is not true (though other grammarians have made the same mistake): a sentence like '`They finally told their adopted son all about himself`' has the pronoun '`himself`' with an indirect object as its antecedent. And (perhaps because of a typing error) example (6.172) on p. 520 is ungrammatical but is not intended to be.

The function of §6.5 is to prepare for §6.6, where Kracht contrasts TG with recent alternative theories, and for §6.7, where he presents some work of his own on formalizing ideas stemming from the TG tradition. The latter is some of the most technical and complex material in the book. It is not adequately prepared for by §6.5's regrettable descent into the kind of loose talk that formalization of linguistic theories is supposed to guard against.

But Kracht was caught between a rock and a hard place: without some kind of introduction to TG his book would be regarded as having slighted mainstream linguistics; yet an introduction that was fully formalized in some way of his own devising would not be recognizable to practitioners. What §6.5 really shows is just how much linguistics needs to improve the rigor of its usual discourse — how much it needs to learn from the rest of this book.

I have not yet required this book for purchase by a class of graduate students. In some ways it would be a fine text, with its rich content and 247 exercises. But I teach on the quarter system, with just ten weeks to a term, and the book's extensive scope makes it very unlikely to be worked through in detail even in a semester. It is also expensive; it does not yet have a paperback edition; the exercises are often very hard, and Kracht provides neither a guide for the instructor nor selected answers to some of the problems for the use of the student. For me, at the moment, it is a personal reference work rather than an every-student-must-buy text.

The criticisms I have made here should be set against the backdrop of my very high opinion of the book as a whole. A reviewer has a duty to draw attention to such faults as can be found, but the ones I have found are mostly minor, and no more than about 30 pages out of more than 600 are affected. Kracht's book will have a place of pride on my mathematical linguistics bookshelf. Looked at in its entirety, it is a remarkable achievement. No other textbook with a linguistic focus (as opposed to a handbook in theoretical computer science) compares with it in topical breadth, technical depth, or awareness of current issues in linguistics. It is the best book of its kind in print today. The future of theoretical linguistics as a serious discipline depends on linguists becoming much better acquainted with the topics Kracht treats here. And certainly, any mathematician, logician, or computer scientist with linguistic interests should invest in a copy.

Acknowledgments This work was partially supported by the Radcliffe Institute for Advanced Study at Harvard University, where the present review was completed. I am very grateful to Gerald Gazdar, Phokion Kolaitis, Osmo Pekonen, James Rogers, Barbara Scholz, Marjorie Senechal, and Stuart Shieber for their careful reading of earlier drafts; several errors were eliminated as a result. Thanks also to Marcus Kracht for generous and helpful responses to queries.

References

[1] AHO, A. V. Indexed grammars — an extension of context-free grammars. *Journal of the Association for Computing Machinery 15* (1968), 647–671.

[2] GAZDAR, G., KLEIN, E., PULLUM, G. K., AND SAG, I. A. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge, MA, 1985.

[3] HAYASHI, T. On derivation trees of indexed grammars — an extension of the uvwxy-theorem. *Publications of the Research Institute for Mathematical Sciences, Kyoto University 9* (1973), 61–92.

[4] IMMERMAN, N. *Descriptive Complexity*. Springer Verlag, Berlin, 1998.

[5] JOSHI, A. K. Tree adjoining grammars: how much context-sensitivity is required to provide reasonable structural descriptions? In *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, D. Dowty, L. Karttunen, and A. Zwicky, Eds. Cambridge University Press, Cambridge, 1985, pp. 206–250.

[6] LEVELT, W. J. M. *Formal Grammars in Linguistics and Psycholinguistics*, 3 volumes ed. Mouton, The Hague, 1974.

[7] MARSH, W., AND PARTEE, B. H. How non-context-free is variable binding? In *Proceedings of the West Coast Conference on Formal Linguistics, Volume Three*, M. Cobler, S. MacKaye, and M. T. Wescoat, Eds. CSLI Publications, Stanford, California, 1984, pp. 179–190.

[8] MONTAGUE, R. *Formal Philosophy*. Yale University Press, New Haven, 1974.

[9] PARTEE, B. H., TER MEULEN, A., AND WALL, R. E. *Mathematical Foundations for Linguistics*. Kluwer Academic, Dordrecht, 1990.

[10] PETERS, P. S., AND RITCHIE, R. W. On restricting the base component of transformational grammars. *Information and Control 18* (1969), 483–501.

[11] POSTAL, P. M., AND PULLUM, G. K. The contraction debate. *Linguistic Inquiry 13* (1982), 122–138.

[12] PULLUM, G. K. The morpholexical nature of to-contraction. *Language 73* (1997), 79–102.

[13] PULLUM, G. K., AND SCHOLZ, B. C. Contrasting applications of logic in natural language syntactic description. In *Proceedings of the 13th International Congress of Logic, Methodology and Philosophy of Science*, P. Hajek, L. Valdes-Villanueva, and D. Westerstahl, Eds. KCL Publications, London, 2005.

[14] ROGERS, J. *A Descriptive Approach to Language-Theoretic Complexity*. CSLI Publications, Stanford, CA, 1998.

[15] ROGERS, J. The descriptive complexity of generalized local sets. In *The Mathematics of Syntactic Structure: Trees and their Logics*, H.-P. Kolb and U. Mönnich, Eds., no. 44 in Studies in Generative Grammar. Mouton de Gruyter, Berlin, 1999, pp. 21–40.

[16] ROUNDS, W. C. Complexity of recognition in intermediate-level languages. In *Proceedings of the 14th Annual IEEE Symposium on Switching and Automata Theory* (Northridge, CA, 1973), IEEE, pp. 145–158.

Radcliffe Institute for Advanced Study
Harvard University
Putnam House, 10 Garden Street
Cambridge, Massachusetts 02138

and after June 30, 2006:

Stevenson College
University of California, Santa Cruz
1156 High Street
Santa Cruz, California 95064-1077

pullum@ucsc.edu