Mathematical linguistics by András Kornai (review)

Emmon Bach, Eric Bach

Language, Volume 89, Number 4, December 2013, pp. 970-973 (Review)

Published by Linguistic Society of America
DOI: 10.1353/lan.2013.0063

Milsark, Gary. 1974. *Existential sentences in English*. Cambridge, MA: MIT dissertation.
Milsark, Gary. 1977. Towards the explanation of certain peculiarities of existential sentences in English. *Linguistic Analysis* 3.1–29.
Verkuyl, Henk J. 1972. *On the compositional nature of the aspects*. (Foundations of language supplementary series 15.) Dordrecht: Reidel.
Verkuyl, Henk J. 1993. *A theory of aspectuality: The interaction between temporal and atemporal structure*. Cambridge: Cambridge University Press.

Department of Linguistics and Department of Spanish, Italian & Portuguese
University of Illinois at Urbana-Champaign
[jonmacd@illinois.edu]

Mathematical linguistics. By András Kornai. London: Springer, 2008. Pp. xiv, 290. ISBN 9781846289866. $99.

Reviewed by Emmon Bach, *University of Massachusetts, Amherst*
and *University of London*,
and Eric Bach, *University of Wisconsin, Madison*

Unlike most books with similar titles, this one is about linguistics and aimed at mathematicians and not the other way around. The preface tells us right away who it is for: 'The book is accessible to anyone with sufficient general mathematical maturity (graduate or advanced undergraduate). No prior knowledge of linguistics or languages is assumed on the part of the reader' (viii).

What is covered? Four traditional linguistic topics—phonology, morphology, syntax, and semantics—are well delineated, with a chapter devoted to each. A fifth, phonetics, is also well represented but scattered through various other sections and chapters. Interleaved with these are four mathematical topics: formal systems for discrete objects, finite-state machines, applications of randomness, and models for continuous phenomena.

In the last chapter, Kornai outlines the themes and challenges of the field (if it is one) and asks: 'Is there, then, a single thread that binds mathematical linguistics together? This book makes the extended argument that there is, and it is the attempt to *find fast algorithms*' (249). Actually, this does not really explain the book's design. A clue to it can be found in K's dissertation (1995). That work is an extended example of explication, in the sense of Carnap (1950), combined with the attempt to bridge the gap between phonological theory and speech engineering, to the potential improvement of both. In *Mathematical linguistics*, K attempts to do the same for theoretical linguistics as a whole. It seems to us, however, that K's presentation reflects the idea of classifying systems by their architectural features, rather than by the running times of algorithms. He also goes well beyond the topics that would be covered in introductory linguistics courses.

K is a phonologist (as well as a mathematician), so a lot of detail and original thinking is evident in the sections on phonology and phonetics. He starts with Bloomfield's postulates (1926) and Pāṇini, and presents the problems and perspectives of phonology from the ground up. Readers with no background in linguistics—supposedly the targeted readers—might do well to examine an introductory linguistics text, lest their eyes glaze over with the free use of articulatory terms such as postalveolar, coronal, and the like. Later parts of the chapter deal with information-theoretic aspects of phonology. Some aspects of phonology are also dealt with in the next chapter (morphology): the prosodic hierarchy, syllables, and stress systems. This chapter deals in depth with the logic of distinctive features, natural classes, and so on.

In the chapter on morphology (Ch. 4), there are no big surprises in K's discussion of word structure and derivational and inflectional morphology. But there are a few mistakes: K misuses the term 'incorporating' for 'polysynthetic' or the like (62), and in discussing the nature of inflections (63), he cites person (in Gunwinggu) as having four values: singular, dual, trial, and plural.

But these are of course values for number, not person. Finally, 'unmarked (i.e. expressed by a zero morpheme)' (63) is surely not right.

Turning to the next chapter (Ch. 5), what syntax is included? As might be expected, K favors frameworks with formally respectable characterizations: dependency grammar, categorial grammar of various persuasions, and the like, but in general the array of theories discussed is very broad, much broader than would be found in typical linguistics textbooks, which tend to make one choice for development in depth. Again, what makes K's exposition unique is the development from the ground up from first principles.

More precisely, K treats three subtopics: combinatorical theories, that is, categorial grammar, phrase structure (§5.1); grammatical theories (making central use of grammatical case), that is, dependency grammars, linking theories, valency (not so much different theories but rather aspects or dimensions of theories) (§5.2); and semantics-driven theories (§5.3). Further sections deal with topics that are usually not treated in linguistic syntax courses and textbooks at all: weighted theories (§5.4) and the 'regular domain' (§5.5), both treated below in some depth.

There are some false claims: that Dutch is V2 in subordinate clauses (98), and that English is an SOV language (103). We also ask again about the intended reader. For example, what will someone with no linguistics make of 'weak crossover', 'heavy NP shift', and 'resumptive pronouns' (76)? These are mostly unindexed and given without examples. (Compare our remarks on phonetic/phonological terminology above.) This chapter can be highly recommended to any syntactician who cares about fundamental principles.

K lays out a number of requirements for semantics (Ch. 6): the system must characterize the set of expressions and the set of meanings; expressions that have similar form should receive analyses by the same apparatus; expressions occurring in fiction should receive analysis by the same apparatus as expressions used in nonfiction contexts; the system must be flexible about naming; and the system must remain invariant under a change of facts.

In a section entitled 'The standard theory', K introduces Montague grammar as a formal theory that 'largely meets, and in some respects goes beyond, the desiderata' just listed (150). The development here follows Richard Montague's 'Universal grammar' (1974, paper 8) rather than the more usual 'The proper treatment of quantification in ordinary English' (PTQ; 1974, paper 7), emphasizing the algebraic nature of the theory: homomorphic relation between the syntactic and the semantic algebras. In exposing Montague's theory, K again goes his own way, wanting to accommodate a more realistic slice of English by taking examples from an actual corpus, rather than the artificially limited 'complete fragment' method of studies that follow(ed) PTQ.

K argues that an account of natural language must allow the expression of possibly contradictory and inconsistent statements. So for a semantic base for the later discussion K takes 'paraconsistent logic' (Priest & Tanaka 2009), which rejects the 'explosive' proposition that from $p \,\&\, \neg p$ everything follows. The last section of the chapter pursues approaches that contend that the semantics 'causes' syntactic form.

There is one problem of detail (153). In presenting Montague's interpretation of noun phrases as generalized quantifiers, K mistakenly renders *a woman* as $\lambda P(\exists x(\text{woman}(x) \Rightarrow P(x)))$ rather than $\lambda P(\exists x(\text{woman}(x) \wedge P(x)))$.

We turn now to the four more properly mathematical topics listed above. K examines three aspects of formal systems: generation, testing, and counting. He presumes that the reader has met these topics before. For example, the simplest generative system is, beyond a doubt, the natural numbers, but in §2.1 K gives us tiling systems, free groups, and Herbrand universes. The usual suspects from the Chomsky hierarchy do make a brief appearance (§2.3), but then K brings optimality theory (§4.3) and categorial grammar (§5.1.2) to the stage. His presentation of the last is motivated by semantics and leaves unanswered the question of how categorial-grammar languages fit into the hierarchy. The distinction between generation and testing is nicely illustrated by balanced parenthesis strings (17), which have both a context-free grammar and a 'checksum' recognition procedure. Finally, one might not imagine counting to be germane to linguistics, but it is, because (as K points out) it tells us how compactly structures can be encoded. For this topic, K's example of choice is the association relations (35) that appear in phonology. (They are related

to the combinatorialist's Delannoy numbers.) Interestingly, the number of these grows exponentially, and K gives us a precise estimate using complex-variable methods.

The basics of finite-state systems are covered very briefly (17–19), and then K moves on to discuss algebraic structures (41–42), tuples of strings (43–49), noncounting languages (133), and languages with weights (§5.4). All of this material will challenge the reader's intuition. For example, the Myhill-Nerode theorem fails to hold for tuples (Regan 1993). (K trips up on this on p. 41.) We recommend this presentation to anyone who thinks finite automata theory stopped with Rabin and Scott.

K connects randomness to linguistics in four ways, by discussing power laws, information theory, statistical inference, and Markov chains. The most familiar power law was used by Zipf and Condon to model word frequencies, but there are many others. K's treatment of them (§4.4) is fascinating and merits close study, but it is not without flaws. In contrast to more rigorous treatments (e.g. Gnedin et al. 2007), K often provides heuristic arguments, which do not tell us whether things happen in expectation, almost surely, or in other ways. (Example 4.4.1 on p. 74 also seems to be wrong, as the number of primes occurring once will be $0(\log N)$.) The discussion of information theory is brief and standard (179–82). Here, K seems to have missed a chance to explain how the search for fast algorithms leads directly to optimal codes, as the naive equal-division approach to Shannon-Fano coding is NP-hard. K's presentation of statistical inference (§§7.2–7.3) highlights concepts from computational learning theory, up to and including the currently favored PAC (probably approximately correct) learning model. He also discusses document classification using word frequencies (§8.4). Perhaps reflecting actual practice, the formal model for this is a bit of a moving target. Markov chains are a probabilistic variation on finite automata, and K's treatment of them (§8.2) emphasizes the popular, but ill-named, hidden Markov model. (What gets hidden is not the model but the state sequence that produced the output.)

Finally, Ch. 9 introduces ideas useful for thinking about how human language gets encoded using continuously varying physical quantities. The theme is Fourier analysis, presented engineering-style. This topic gives a bridge between the continuous and discrete worlds, and it is tremendously important technologically. K motivates this by explaining that, historically, almost all of the computation in speech recognition systems was just signal manipulation. Accordingly, we get a quick review of some of the popular coding methods from digital signal processing: pulse code modulation, nonlinear signal warping (usually called companding, and related to conjugation in dynamical systems), delta modulation, and the current favorite, linear predictive coding. The descriptions are clear and to the point, and should be read by all linguists who want to communicate with engineers.

We now consider the book as a whole. Despite claims to the contrary, the book is not a textbook, as its style is more similar to a long expository article. In particular, many of the exercises are really small (or not so small) research topics. K is honest about this (ix).

The table of contents sufficiently guides us to the book's major topics, but ideas can also vanish and reappear much like streams from an underground river. Occasionally the distance between a concept's first mention and its definition is quite wide. This can be problematic, since the numbering scheme for theorems and related statements (probably from LaTeX) makes it hard for readers to get around. Thus, the most useful improvement to the book would be a better index. (Readers, sharpen your pencils!)

The scope of the book is enormous, and K deserves a lot of credit for pulling together so much disparate material into fewer than 300 pages. There are occasional rough spots in the exposition (e.g. the cumbersome formal definitions for association relations), but these can be easily remedied by those familiar with the computer scientist's approach to discrete structures. Except for a few errors, which we have tried to note above, the linguistic and mathematical technicalities are accurate.

Who would use this book, and how? The ideal reader is probably a brilliant student who has learned something of mathematics, theoretical computer science, linguistics, and information theory and wants to see how it all fits together. Professionals will also want it on the shelf for quick introductions to topics they do not know. Those using it for courses should choose a few topics and supplement the book with some classic papers.

Since the scope of the book is vast, readers should heed K's advice about what is needed to work through it. But linguists should also be encouraged to tackle it as a way of learning about a lot of material that is inherently fascinating and might give them totally new perspectives on what they do for a living. In short, the book really delivers on its promise, and presents topics old and new that lend themselves very well to thinking mathematically about language.

REFERENCES

Bloomfield, Leonard. 1926. A set of postulates for the science of language. *Language* 2.153–64.
Carnap, Rudolf. 1950. *Logical foundations of probability*. Chicago: University of Chicago Press.
Gnedin, Alexander; Ben Hansen; and Jim Pitman. 2007. Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probability Surveys* 4.146–71.
Kornai, András. 1995. *Formal phonology*. New York: Garland.
Montague, Richard. 1974. *Formal philosophy: Selected papers of Richard Montague*, ed. by Richmond H. Thomason. New Haven, CT: Yale University Press.
Priest, Graham, and Koji Tanaka. 2009. Paraconsistent logic. *Stanford encyclopedia of philosophy*. Online: http://plato.stanford.edu/entries/logic-paraconsistent/.
Regan, Kenneth. 1993. Formal language theory: Origins and new directions (Class notes for CS682: Formal Language Theory II). Buffalo, NY: Department of Computer Science, SUNY Buffalo. Online: ftp://ftp.cse.buffalo.edu/users/regan/LNall2page.ps.

Emmon Bach
University of Massachusetts, Amherst and
The School of Oriental and African Studies, University of London
[ebach@linguist.umass.edu]

Eric Bach
Computer Sciences Department
University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706
[bach@cs.wisc.edu]

A grammar of Saramaccan Creole. By John C. McWhorter and Jeff Good. (Mouton grammar library 56.) Berlin: De Gruyter Mouton, 2012. Pp. xi, 246. ISBN 9783110276435. $210 (Hb).

Reviewed by Peter Bakker, *Aarhus University*

De Gruyter Mouton must be praised for the series of excellent grammars, which was started almost thirty years ago. Almost sixty grammars have been published, and this grammar of Saramaccan Creole is the least voluminous. All other grammars cover twice to almost four times as many pages. Does that mean that the description of Saramaccan requires fewer pages? And is this a significant observation in the light of creole studies?

Saramaccan is an English-lexifier creole spoken by some 50,000 people, most of whom live in Suriname, but a significant number have also settled in neighboring French Guyana since the 1980s, and some families are in the Netherlands and the US. Saramaccan is spoken by descendants of maroons whose communities in interior Suriname were formed between the 1690s, when the first slaves escaped and settled in the rainforest, and 1762, when the maroons agreed not to accept any more refugees in a peace treaty with the Dutch. The language also shows quite a bit of impact from Portuguese in the lexicon, as some escaped slaves came from plantations run by Portuguese-speaking Sephardic Jews. Saramaccan is the creole language that developed with the least influence from the lexifier, and therefore it may shed light on language creation more than any other new language, as it developed in the absence of a model or target language. Fewer than a few thousand roots from English, Portuguese, and African languages were available to create a