The original of this review appeared in the last ever print issue of CL, December 2008. I've added some comments and clarifications – these are typeset in blue, often as footnotes (there are no footnotes in the original). András Kornai, February 2009

Mathematical Linguistics

András Kornai
MetaCarta Department of Algebra, Budapest Institute of Technology

London: Springer-Verlag (Advanced Information and Knowledge Processing series, edited by Lakhmi Jain and Xindong Wu), 2008, xiii+289 pp; hardbound, ISBN 978-1-84628-985-9, eISBN 978-1-84628-986-6, $99.00. Originally, currently $79.11 on Amazon.

Reviewed by
Richard Sproat and Roxana Gîrju
University of Illinois at Urbana-Champaign

For readers of traditional textbooks such as (Partee, Ter Meulen, and Wall, 1990), the term 'mathematical linguistics' denotes a rather narrowly circumscribed set of issues including automata theory, set theory, lambda calculus with maybe a little formal language theory thrown in. Kornai's contribution is refreshingly different in that he treats, in this relatively compact volume, practically all areas of linguistics, phonetics and speech and language processing.

Kornai's motivation for writing this book is to present "a single entry point to the central methods and concepts of linguistics that are made largely inaccessible to the mathematician, computer scientist, or engineer by the surprisingly adversarial style of argumentation . . . and the proliferation of unmotivated notation and formalism . . . all too often encountered in research papers and monographs in the humanities." There is no question that much of what passes for rigor (mathematical *and* scientific) in linguistics is a joke, and that there is clearly a need for any work that can place the field on a more solid footing. It also seems likely that Kornai is the only person who could have written this book.

The book is divided into ten chapters, including a short introductory chapter, which lays the groundwork and identifies the potential audience, and a concluding chapter where Kornai reveals his own views on what is important in the field, which in the interests of balance he has largely suppressed throughout the book. Chapter 2 is also introductory in that it presents basic concepts of generation (via a ruleset), axioms and string rewriting.

The main chapters (3–9) deal with a variety of topic areas relating to language and speech, starting with phonology in Chapter 3. This chapter introduces the notion of phonemes, distinctive features, autosegmental phonology and computation using finite automata. Kornai offers many details that are of course lacking in most linguistic treatments, such as a proof that the number of well-formed association lines between two tiers of length n is asymptotically $(6 + 4\sqrt{2})^n$. This is Theorem 3.3.1 on p. 36 – see the discussion on p. 37 why this matters.

Chapter 4 deals with morphology, which for Kornai includes not only word formation, but also prosody (including stress assignment and moraic structure), as well as Optimality Theory and Zipf's law.

The fifth chapter treats syntax, including categorial grammar, phrase structure, dependency frameworks, valency and weighted models of grammar, ending with a discussion of weighted finite automata and hidden Markov models. In the context of weighted models Kornai implies that Chomsky's original notion of degree of grammaticality fits naturally as an instance of a weighted model with a particular semiring; of course, exactly what the $\oplus$ and $\otimes$ operators of that semiring map to remain to be seen insofar as the notion 'degree of grammaticality' has never been rigorously defined, a point Kornai should have made himself.

Chapter 6, on Semantics, starts with a discussion of various standard paradoxes such as the Liar, and then moves on to an overview of Montague's theory, type theory and grammatical semantics. Throughout the discussion, Kornai underscores the fundamental limitations of theories of semantics that are based purely upon evaluation of truth conditions for artificial fragments, an important point for anyone who wants to go beyond theoretical philosophically inspired models and consider semantic interpretation in the real world.

Complexity is the topic of Chapter 7. This is not the Chomsky-hierarchy notion of complexity, but rather deals with information theory, in particular entropy, Kolmogrov complexity, and a short section on learning, including identification in the limit and PAC learning.

Pattern recognition is divided across two chapters, with Chapter 8 laying the essential groundwork of linguistic pattern recognition, and Chapter 9 presenting details on speech processing and handwriting recognition. This includes feature extraction: in the case of speech recognition, Kornai reviews the frequency representation of speech signals, and defines the cepstrum. Discussion of acoustic models leads us to phonemes as hidden units, with a slight detour into the fine-grained distinctions between different levels of phonemic analysis in the once popular but now largely discredited theory of Lexical Phonology.[1]

Each chapter ends with a section entitled "Further Reading", and the texts referred to are generally quite useful as material for readers who wish to explore the issues further.

According to Wikipedia, Kornai is a "well-known mathematical linguist", whose Erdős number is 2.[2]Unfortunately, neither of us can claim Kornai's mathematical sophistication or stature, but on the other hand, this makes us good judges of the book's potential audience; and herein lies a problem. Kornai's target is "anyone with sufficient general mathematical maturity" with "[n]o prior knowledge of linguistics or languages ... assumed on the part of the reader" (p. viii). This suggests that the book is not primarily aimed at linguists, and certainly the mathematical maturity assumed puts this book well beyond the reach of most linguists, so that it could not easily be used in an introductory course on mathematical linguistics in a linguistics program. It is probably beyond the reach of many computer science students as well.

What about those who do have the mathematical maturity, but know nothing about linguistics? The problem here is that in many cases Kornai does not give enough background (or any background) to appreciate the significance of the particular issues being discussed. For example, on p. 77 Kornai gives *weak crossover* and *heavy NP shift* as examples of phenomena that have 'weak' affects on grammaticality, and *resumptive pronouns* as examples of phenomena that are marginal in some languages (such as English).

> In most, though not necessarily all, cases it is relatively easy to construct pairs of sentences, one grammatical and the other not, that bring into sharp relief how a particular rule or constraint operates or fails to operate. There are areas of grammar such as 'weak crossover' or 'heavy NP shift' which are *weak* in the sense that the contrast is less visible and obvious than in the examples above, but even if we raise the bar very high, there are plenty of significant contrasts left for a theory of syntax to account for. On the whole, the development of syntax is not crucially impacted by the weaker examples, especially as there are generally other languages where phenomena marginal in one language, such as resumptive pronouns in English, can be observed in unequivocal examples, and often in far richer detail.

1 For an opposing viewpoint, the reader may want to consult Rubach (2008)
2 Wikipedia neglects to mention this, but Kornai's Kibo number is 1.

But nowhere does he explain what these terms denote, which means that these are throw-away comments for anyone who does not already know.[3]Section 3.2 introduces phonological features and feature geometry and sketches some of the mathematical properties of systems with features; but very little background is given on *what* features are supposed to represent. The short discussion of Optimality Theory (pp 67–69) hardly gives enough background to give a feel for the main points of that approach. In other cases, topics are introduced but their importance to surrounding topics is hard to fathom. For example, in 6.1.3 a discussion of the Berry paradox leads into a digression on how to implement digit-sequence-to-number-name mappings as finite state transducers. Apart from giving Kornai an opportunity to emphasize that this is trivial to do (something that is true in principle, but less true in practice, depending upon the language), it is not clear what purpose this digression serves.

> Aside from the tradition of Indian logic, the study of languages had very little impact on the foundations of mathematics. Rather, mathematicians realized early on that natural language is a complex and in many ways unreliable construct and created their own simplified language of formulas and the mathematical techniques to investigate it. As we shall see, some of these techniques are general enough to cover essential facets of natural languages, while others scale much more poorly.
>
> There is an interesting residue of foundational work in the Berry, Richard, Liar, and other paradoxes, which are often viewed as diagnostic of the vagueness, ambiguity, or even 'paradoxical nature' of natural language. Since the goal is to develop a mathematical theory of language, sooner or later we must define English in a formal system. Once this is done, the buck stops there, and questions like "what is the smallest integer not nameable in ten words?" need to be addressed anew.

There are also a number of places where issues are presented in a non-standard way, which might make sense from some points of view, but not if you are trying to introduce someone to the way the field is practiced. It is odd, for instance, that prosody is introduced not in the chapter on phonology, but in the one on morphology. It is also somewhat odd that Zipf's law gets introduced in the morphology chapter. (And why is it that nowhere does Kornai cite Baayen's excellent book on word frequency distributions (Baayen, 2001), which would be a very useful source of further information on this topic to any reader of Kornai's book?)[4]

Some material presented is puzzling or simply wrong. It is not explained in what sense German has a "pure SVO construction" (p. 103) in contradistinction to the normal assumption that German is verb second.[5]The Cypriot syllabary does *not* date from the 15th century BCE (p. 54)[6]; Latin does *not* have two locative cases (p. 90) — indeed, it does not even have one locative case, so-called[7]; the basic Hangul letter shapes (introduced on p. 31 to make a point about phonetic features)

3 Clearly, the intent of the quoted passage is debatable. Many, if not most, readers of CL would feel that historically the development of syntax has been crucially impacted or even distorted by excessive reliance on flimsy made-up examples and statistically negligible pseudo-phenomena. The immediate concern raised by the reviewers would be trivial to fix by adding a few references such as Wasow (1979) to the original discussion. But keep in mind that this whole passage occurs in the introduction to the syntax chapter. The literature on these specific subjects is immense, and much of it would simply alienate, rather than educate, the empirically-minded reader.

4 Because he finds the volume in question less than excellent, see Kornai and Penn (2003).

5 Sentences like *Der Hund biss den Jungen* are neither strange nor rare.

6 True but perhaps irrelevant given that the passage in question is about the time depth of syllabic/moraic writing, and Cyprio-Minoan is conventionally dated that far.

7 The Latin locative is marginal, to be sure, but dismissing its existence out of hand seems a bit rash.

> No single language has phonemes at every point of articulation, but many show five-, or six-way contrasts. For example, Korean distinguishes bilabial, dental, alveolar, velar, and glottal, and the difference is noted in the basic letter shape (□, ∨, ⌐, ⌐, and ◯, respectively).

are, with two exceptions, completely incorrect — probably it would have been better to use a real Korean font rather than trying imitate the *jamo* with LaTeX math symbols.[8] There are of course a great many places where the discussion is useful and informative, but there are enough examples of the kinds we have outlined that the uninitiated reader should be careful.

As far as we can see, the most likely readership of this book consists of (computational) linguists and others who already know the linguistic issues, have a fairly strong formal and mathematical background, and could benefit from the more precise and more rigorous mathematical expositions that Kornai provides.

Throughout the book, Kornai pauses occasionally to present exercises to the reader. These range from relatively simple, to major research projects. As with other aspects of this book, the distribution of topics for the exercises is somewhat erratic. Thus, on p. 184, in the chapter on complexity, we are offered exercises 7.6 and 7.7 in close proximity:

> **Exercise 7.6** Prove that a regular language is prefix-free iff it is accepted by a DFSA with no transitions out of accepting states. Is a prefix-free language context-free iff it is accepted by a DPDA with the same restriction on its control?
>
> ⋮
>
> **Exercise 7.7** Research the role of the ascii codes 0x02 (STX), 0x03 (ETX), and 0x16 (SYN).

But variety is, after all, what keeps things interesting.

References

Baayen, R. H. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Kornai, András and Gerald Penn. 2003. Vocabulary: imitative or generative? ms.

Rubach, Jerzy. 2008. An overview of lexical phonology. *Language and Linguistics Compass* **2** 456-477.

Partee, Barbara, Alice Ter Meulen, and Robert Wall. 1990. *Mathematical Methods in Linguistics*. Springer, Boston, MA.

Thomas Wasow. 1979. *Anaphora in Generative Grammar*. Story-Scienta, Ghent

Richard Sproat is Professor of Linguistics and Electrical & Computer Engineering at the University of Illinois, Urbana Champaign. He works on computational morphology, text normalization and speech processing. His Erdős number is 4.

Roxana Gîrju is Assistant Professor of Linguistics at the University of Illinois, Urbana Champaign. She has a PhD degree in Computer Science and works on computational semantics, pragmatics and inference. Her Erdős number is also 4.

8 The shapes were taken from pedagogically oriented works, which offer the right level of simplification for the point being made.