Mach Translat (2007) 21:253–256
DOI 10.1007/s10590-008-9043-4

BOOK REVIEW

Kornai, András: Mathematical Linguistics

Springer, London, 2008, xiv + 290 pp

Horacio Rodríguez

Received: 3 October 2008 / Accepted: 9 October 2008 / Published online: 2 November 2008

I have read this book with pleasure and the first adjective I have to choose to describe it is "fascinating". The book is very well written, the author has a wide background, not only restricted to mathematics and linguistics, and bits of it are spread along the book in form of original settings, well-motivated examples, interesting proposed exercises and highly elaborated "further reading" sections at the end of each chapter. The presentation, discussion and, in some cases, solution of some paradoxes of natural language are brilliant examples of its interesting content.

The book has, however, some drawbacks. The most important is that it is not clear at all who is the intended audience of the book. The publisher (Springer) says that "the book is addressed to computer scientists, engineers and mathematicians interested in NLP [natural language processing]", but later "essential reading for every linguist". The author claims that the "the book is accessible to anyone with sufficient general mathematical maturity (graduate or advanced undergraduate) [...] No prior knowledge of linguistics is assumed". In another point the author acknowledges to be more interested in competence than in performance. This is an obvious warning against readers interested in empirically based NLP. In my opinion a rather high level of background knowledge in mathematics is needed and also a reasonable, although perhaps not so high, level of linguistic knowledge is required. So the intended audience is rather narrow.

Another drawback is that the content of the book is somewhat irregular in several aspects.

The first irregularity affects the organization of the book. After two introductory chapters, the core of the book is organized according to the standard layer-based

H. Rodríguez (✉)
Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
e-mail: horacio@lsi.upc.edu

structure of linguistic phenomena: phonology, morphology, syntax and semantics. This is a reasonable approach. The rest of the book contains, however, other complementary material (i.e. chapters on complexity, pattern recognition, speech & handwriting and simplicity) that has not been organized in a systematic way (e.g. in a dimension somewhat orthogonal to the basic structure, or including new layers into it, for instance, adding a phonetic layer to the four described in the book). I have the impression that these chapters have been included in the book without insuring sufficient cohesion with the core chapters of the book.

The second source of irregularity regards the informative content of the book. There are clear differences in quality, coherence of the description and coverage between the different chapters. Chapter 2 presents the basic mathematical elements over which the linguistic constructs will be represented. Chapters on phonology and morphology, by far the most interesting (unsurprisingly, considering Kornai's profile), follow with fidelity the representational issues proposed in Chap. 2. The remaining chapters depart considerably from this mathematical base.

A final source of irregularity refers to the level of details in explanations which, in some cases, correlates poorly with the importance of the topic. For instance, Theorem 3.3.1 states that the number of well-formed association relations over two tiers, each containing a string of length n, is asymptotically $(6+4\sqrt{2})^n$. It is a moderately important theorem but Kornai devotes one page to the proof. Other theorems, in my opinion more important, such as the four included in Sect. 2.3 relating to context-free and context-sensitive languages and grammars, are presented without proof, with only short comments, or with no comment at all.

It is clear that Kornai is a big fan of Pānini. Acknowledging the importance of Pānini's work, especially of the Ashtadhyayi, is fair and constitutes a lesson in humility for those who believe that all the important issues on computational or formal linguistics started with Chomsky, Harris or de Saussure, but in this book Pānini becomes omnipresent and I have found this somewhat exaggerated.

As mentioned above, the book is organized in 10 chapters. Each chapter begins with a brief introduction serving as a reading guide for the chapter which relates its content to that of the other chapters and ends with an extremely interesting "further reading" section. Most chapters contain a lot of interesting exercises, and in general the exercises are very challenging and need further reading in order to be tackled. An extreme example of such challenges is Exercise 4.2: "Develop rules describing the placement of accents in Sanskrit", where the reader needs to know not only mathematics and linguistics but also Sanskrit to solve it!

I will continue with some details regarding the core of the book.

Chapter 3 deals with phonology. The chapter presents in a really brilliant form the concepts of phoneme, natural classes (I found the discussion of this issue particularly interesting), tiers, bistrings, bilanguages and so on. The presentation of finite-state automata (FSAs) and finite-state transducers (FSTs) and many variants of these machines is simply correct but I have seen many better presentations than this one (see, for instance, the well known book "Finite-State Language Processing", edited by Emmanuel Roche and Yves Schabes, MIT Press, 1997) . The usefulness of FSAs and FSTs for many NLP tasks is well known to NLP practitioners but is not evident for many other potential readers of this book. As a result of this, I found the treatment of finite-state formalisms rather poor. Perhaps this small disappointment comes from the high expectations I put on this topic, the closest to Kornai's profile.

Chapter 4 is devoted to morphology. The prosodic hierarchy and their components, syllables, moras, feet, and so on, are adequately presented and motivated. The presentation of word formation, introducing the concepts of morpheme, root, affix, morphological paradigm, etc. is well explained and accurate. I have found, however, an important omission: the important topic of morphology learning (morpheme discovery, morphological rule learning and so on) is neither described nor referenced. Important mathematical techniques, such as minimum description length (MDL) could have been introduced in a similar way. A brilliant contribution of this chapter is the presentation of Zipf's law and related issues.

Chapter 5 is devoted to syntax. Kornai proposes three approaches to syntax:

- Combinatorial, i.e. based on the relations between syntactic components (syntagms) and their neighbours.
- Grammatical, i.e. based on the internal structure of the syntactic components and
- Semantic, based on the fit between the utterance (what is said) and what is seen in the world.

In this presentation Kornai places categorial and phrase-structure grammars in the first approach and dependency grammars in the second, while placing the so-called semantic-driven theories in the third approach. I think it is an interesting and novel approach to syntax.

He pays special attention to the formalisms whose expressivity can be placed between context-free languages (CFL) and context-sensitive languages (CSL), as many natural-language phenomena can be described in this form. I found this discussion very interesting; however, I must point out two limitations:

- The description of unification-based syntactic approaches is far from being complete. Both free (as Patr-II, generalized phrase structure grammars (GPSG), lexical functional grammars (LFG), etc.) and type-constrained (as head-driven phrase structure grammars (HPSG) at linguistic level or attribute logic engine (ALE), comprehensive unification formalism (CUF), typed feature structure (TFS), etc. at computational level) approaches should be presented for the sake of completeness. The case of type logic formalisms, reduced to a simple reference, is paradigmatic. Type logic formalisms provide a theoretical support for both type-based syntactic theories and constraint-based languages, as ALE.
- The description of probabilistic approaches is good (weighted theories, weighted rules, etc.) but incomplete. Many novel results on probabilistic parsing (generative versus discriminative models, ranking approaches, etc.) should have been included in this chapter.

The presentation of probabilistic versions of previous models, as weighted finite-sate automata, (WFSA), weighted finite-state transducers (WFST) or probabilistic context-free grammars (PCFG) is quite accurate; the presentation of hidden Markov models (HMMs), however, is incomplete. HMMs are undoubtedly a very useful resource but currently it has to be placed in the context of more expressive mechanisms (such as, for instance, conditional random fields). Presenting HMMs in the context of graphical models, with their subclasses directed and undirected graphical models, etc. might have been a good option. I found the section on external evidence, which analyses the complexity of parsing very interesting.

Chapter 6 deals with semantics. The chapter presents first the well known formalism by Montague and then a more performance-oriented approach based on a semantic calculus. This new approach is especially appealing.

The rest of the book also contains valuable material but, as I said before, it is included in a rather unsystematic way. The presentation of concepts in information theory (entropy, mutual information, etc.) and Kolmogorov complexity is good. The section devoted to machine learning is rather superficial although it contains a good description of MDL. Chapter 8 once more presents the HMM machinery and devotes a section to document classification. The description is good but I wonder why document classification is the only natural-language (NL) application described in the book. The final part of the book presents some material on speech and handwriting.

An omission has to be pointed out that can be of particular interest for the readers of this journal. This book provides clear and useful insights into the mathematics of natural languages. No particular language is addressed and examples cover very different ones. No attempt is made, however, to provide mathematical basis for the mappings between languages. Including this material could be useful as mathematical basis for many multilingual NL processing tasks, including, obviously machine translation.

A final note is needed. When I was asked to read and comment on this book I went back to the good old book by Barbara Partee, Alice ter Meulen and Robert E. Wall on the same subject ("Mathematical Methods in Linguistics", Kluwer Academic, 1990), which has been an indispensable reference for years. A comparison is needed to conclude this report. The intended audience of both books is very different, as well as the organization and the point of view. Partee's is addressed to linguists lacking a background in mathematics and provides only the mathematical material needed for managing the linguistic concepts. Kornai's book needs a higher level of mathematical background in order to benefit from and enjoy it. In some way, the books are complementary.