Probabilistic Linguistics

Rens Bod, Jennifer Hay, and Stefanie Jannedy (editors)
(University of Amsterdam, University of Canterbury, and Lucent Technologies)


Reviewed by
Patrick Juola
Duquesne University

For nearly 40 years, the generally accepted view among linguists has been that “language” is a categorial structure, with probability relegated to perhaps explaining errors of human “performance.” In the past 10 years, theoretical developments, including example- and statistical-based parsing, the mature development of connectionist natural language processing, and stochastic versions of traditional theories (such as stochastic optimality theory), have amassed much evidence that this view is limiting and that replacing categorial structures with probabilistic reasoning will result in better performance.

A specific countercriticism of the quantitative practices in natural language processing is that the kinds of reasoning employed are unnatural and in many cases ignore important observations about linguistics; examples of this include the use of Markov processes or n-gram statistics to describe phonology or syntax, instead of the more linguistically plausible (and traditional) grammar-based formalisms. In this book, Bod, Hay, and Jannedy present some of the evidence in favor of a role for probability in linguistic reasoning. Moreover, they also argue that incorporating probability theory into linguistic theory is not only legitimate but necessary. Probability theory should be compatible with, instead of a substitute for, more traditional linguistic observations.

The book itself is an outgrowth of a symposium organized by the editors at the Linguistic Society of America 2001 annual meeting held in Washington, D.C. Much of the book is thus dedicated to presenting background material specifically for non-computational linguists, who might not have a strong mathematical and statistical background. Chapter 2, an introduction to the theory of probability that starts more or less at the very beginning, is thus either entirely useless or the most important chapter in the book, depending upon the reader’s background. Fortunately, it is well organized, well written (for those that need it), and easily skippable (for those that don’t).

The rest of the book is taken up by individual chapters by noted experts in the various subfields of psycholinguistics, sociolinguistics, historical linguistics and language change, phonology, morphology, syntax, and semantics. The experts are well-chosen and include leading lights such as Baayen, Jurafsky, Manning, and Pierrehumbert, as well as the editors themselves. Most of the individual chapters are similarly structured; the author(s) observe that categorial theories do not neatly account for observed variation and then present a probabilistic model that accounts for both categorial and the observed continuous data. These models are fairly specific to the problems studied by each expert, and together they make an interesting collection of different ways to solve linguistic problems from a variety of standpoints; students of modeling could do much worse than to simply page through and look at each model in turn to see whether it could be adapted to their own studies.
From a broader perspective, a key aspect of the models is that they better account for the observed behavior of real people and real languages. Although “frequency” has long been known to be a very important variable in much of psychology and in psycholinguistics in particular, these models provide specific mechanisms that illustrate how frequency can produce the known effects. Examples of this include Pierrehumbert’s discussion of category discriminability among phonemes, Jurafsky’s discussion of final-consonant deletion and latency, Mendoza-Denton et al.’s discussion of the sociophonetics of /ay/, and Baayen’s discussion of Dutch linking elements. In each case, the effects of frequency are directly modeled in a probabilistic framework and an appropriate causal role is assigned.

One specific point of interest from a more theoretical viewpoint is Jurafsky’s discussion of the psychological reality of probabilistic models (“Surely you don’t believe that people have little symbolic Bayesian equations in their heads?”). He addresses many of the challenges that traditional linguistics might present to probabilists. The bibliography is extensive, and there is a useful glossary of probabilistic terms to help readers keep definitions in mind.

From the standpoint of computational linguistics, the book is slightly disappointing in not discussing computational processes more, as the reader is (usually) left to infer the exact mechanisms and algorithms used to implement the equations discussed in the book. Another weakness is the lack of discussion of statistical inferences, which are often necessary to interpret the probabilistic models themselves (is it reasonable, for example, to expect readers who actually need chapter 2 to understand generalized linear modeling?). The most significant omission is a discussion of how these individual models interact, either with each other or with more traditional categorial models of other linguistic subfields. These are minor weaknesses in an otherwise significant work. For people who already believe that “probabilistic linguistics” is a sensible and meaningful phrase, this book may be useful as a catalog of how probability theory can be and has been applied to the various linguistic subdisciplines. For those who hold to categorial theories of language, this book will at least provide a single source for some major arguments, evidence, and theories supporting probabilistic processes underlying linguistic competence. And for those who believe that probability is really only a description (a quantification of our ignorance, if you will), this volume neatly summarizes ways in which probability may play a key role in human language processing.

Patrick Juola has been working in computational and statistical applications of psycholinguistics since the early 1990s as a Ph.D. student at the University of Colorado. He is currently an assistant professor of computer science at Duquesne University, 600 Forbes Avenue, Pittsburgh, PA 15282; e-mail: juola@mathcs.duq.edu.