Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax

Emily M. Bender
(University of Washington)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 20), 2013, xvii+166 pp; paperbound, ISBN 978-1-62705-011-1, $40.00; e-book, ISBN 978-1-62705-012-8, $30.00 or by subscription

Reviewed by
Chris Dyer
Carnegie Mellon University

The phenomenal success of machine learning in engineering natural language applications has led to a curious situation: Natural language processing practitioners who were trained in the last 15 to 20 years may have established a quite successful career in this area with only a haphazard knowledge of the science of natural languages. The premise of the new volume by Emily M. Bender is that greater awareness of linguistics will enable continued technical progress, particularly as language applications are required to perform more intelligent processing in more languages.

This book is not beholden to any particular theoretical program. Rather, it is a survey of the morphological and syntactic means by which different languages express meaning, anchored by clear and effective examples from typologically diverse languages. Eschewing theorizing to stay close to data permits a remarkably wide range of linguistic phenomena to be covered, and it is this that is the book's greatest strength. However, in a few places, a seemingly arbitrary theoretical perspective is assumed rather more tacitly than one might hope, with few hints as to alternative analyses (e.g., see the following remarks about parts of speech in Chapter 6). Furthermore, a bit more theoretical scaffolding could have made the presentation more succinct in places (e.g., Chapter 7's excellent discussion of heads, arguments, and adjuncts could have been more precise with a basic logical calculus). Finally, although theoretical squabbles can be off-putting to outsiders, theoretical diversity can have practical benefits, particularly in a field as omnivorous as NLP. For example, while theorists might disagree about whether morphophonology is best modeled with systems of rewrite rules (e.g., SPE) or constraint satisfaction (e.g., Optimality Theory) (Chomsky and Halle 1968; Prince and Smolensky 2004), each suggests a distinct computational instantiation with different challenges and opportunities. For such reasons, more discussion of theory would not have been unwelcome. This slight objection aside, the book is an excellent introduction to the diversity of linguistic representations that NLP must eventually contend with.

The book is organized into 10 chapters, in roughly two parts (the first part, morphology; the second, syntax), spread over 100 numbered topics.

Chapter 1 gives an overview of the scope of the book, distinguishing morphology and syntax from bag-of-words models. It lays out the premise that knowledge of linguistic structure can guide engineers in profitable directions by facilitating error

doi:10.1162/COLI_r_00212

analysis and feature engineering. The notion of bounded variation is introduced: the idea that while languages exhibit diversity in how they pair sound and meaning, this variation is subject to limits, and that different languages can have similarities due to areal, genetic, and typological relatedness. A brief survey of the genetic taxonomy of the world's languages is given and the number of speakers they have—as well as the striking difference in distributions of the languages in the NLP literature.

Chapters 2 and 3 introduce morphology and morphophonology, focusing on the internal structure of words and how they are realized in text and speech. Simple English examples motivate the discussion, but more exotic nonconcatenative processes in Semitic languages and infixation examples from Lakhota emphasize phenomena that may be unfamiliar to those with experience only with Indo-European languages. The conventional tripartite distinction of roots and derivational and inflectional affixes is presented to organize the kinds of meaning/function changes characteristic of morphological processes, although compounding and cliticization—which fit less neatly into this taxonomy—are also discussed. Because syntax and semantics were only briefly mentioned in the Introduction, the extensive forward references to the related material in the later chapters were quite helpful for clarifying terms, making the e-book version particularly convenient.

Chapter 4 discusses morphosyntax, reviewing the diverse grammatical functions that different languages encode with morphology. Phenomena covered include functions applying to the verbal domain, including tense, mood, and aspect, negation, evidentiality; the nominal domain, including person, number, and gender, case, definiteness, and possession; and various common agreement processes.

Having established how words are constructed from morphemes, Chapters 5–9 focus on how syntax is used to combine words to form an unbounded number of sentences whose meaning is determined compositionally. Chapter 5 introduces the distinction between grammaticality of sentences and how syntactic structure determines their meanings, and Chapter 6 introduces parts of speech as clusters of distributional regularities of words and phrases in grammatical sentences. The fact that discussion of grammaticality proceeds almost exclusively in terms of POS—a familiar construct to anyone working in NLP, but one that looks quite different in many theories of syntax (Steedman 2000; Stabler 1997)—is a shortcoming.

Chapter 7, perhaps the strongest in the book, discusses syntax in terms of headed phrases that relate to each other either as arguments (which semantically complete the meaning of a predicate) or adjunction (which introduces additional predicates). Diagnostics for distinguishing heads and dependents as well as arguments and adjuncts are given, together with clear examples of their application, and common mistakes (e.g., using optionality as a test of argumenthood or assuming that only verbs can select arguments) are covered. A particularly useful part of this chapter is a discussion of lexical resources (FrameNet, ProbBank) and how they relate to the concepts being discussed.

Chapter 8 discusses argument types and grammatical functions, reviewing not entirely successful attempts to create universal inventories of thematic roles, ultimately demonstrating that syntactic roles are less idiosyncratic (at least within single languages), and capture many generalizations useful for semantic analysis. A discussion of cross-linguistic properties of subjects and the distinction between core and oblique arguments follows. Three important sections discuss the subtle and often confusing distinctions between syntactic and semantic arguments with effective examples. Although most of this chapter focuses on English examples, various morphological strategies for marking grammatical functions is discussed.

Chapter 9 concludes the syntax portion of the book, focusing on the processes that can introduce divergences between syntactic and semantic relationships. Because such divergences underlie many constructions with considerable value in NLP (e.g., wh-questions in English) and directly challenge the simplifying assumption of transparency between syntax and semantics, it is fortunate that this section goes into considerable detail, covering phenomena including passivization, dative shift, expletives, raising, control, and various kinds of long distance movement, as well as a good discussion of phenomena found in other languages, such as causative morphology and discontinuous constituents.

Chapter 10 provides a brief appendix summarizing various large-scale computational resources (morphological analyzers, parsers, typological atlases) that encode linguistic knowledge.

This book serves as a useful introduction to linguistic phenomena that will help NLP researchers orient themselves with respect to phenomena they will encounter as their applications push into new languages and strive for deeper automated understanding of language. The tension between the science of linguistics and natural language engineering and the resulting missed opportunities has been remarked upon in these pages recently (Wintner 2009), and we should applaud this successful effort to find common ground.

References

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row.

Prince, Alan and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Wiley–Blackwell.

Stabler, Edward P. 1997. Derivational minimalism. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*, volume 1328 of *LNCS*. Springer, Berlin Heidelberg, pages 68–95.

Steedman, Mark. 2000. *The Syntactic Process*. MIT Press.

Wintner, Shuly. 2009. Last words: What science underlies natural language engineering? *Computational Linguistics*, 35(4):641–644.

This book review was edited by Pierre Isabelle.

Chris Dyer is an assistant professor in the Language Technologies Institute and Machine Learning Department at Carnegie Mellon University. His primary research focus is the application of machine learning techniques to problems in multilingual natural language processing. Together with Jimmy Lin, he is co-author of *Data-Intensive Text Processing with MapReduce*, published by Morgan & Claypool in 2010. Dyer's e-mail address is cdyer@cs.cmu.edu.