Empirical Linguistics

Geoffrey Sampson
(University of Sussex)

London: Continuum (Open linguistics series, edited by Robin Fawcett), 2001, viii+226 pp; hardbound, ISBN 0-8264-4883-6, $90.00, £55.00

Reviewed by
Steven Abney
AT&T Laboratories–Research and University of Michigan

Geoffrey Sampson has made significant contributions in the area of corpus linguistics, and this book brings together and updates a number of his essays, mostly from recent years but including work whose original publication dates back as far as 1975. Despite this variety of provenance, the volume has been well edited for consistency both in theme and in style. Sampson's stated aim is to give a coherent presentation of an approach to language that he has enunciated in scattered publications over the years, an approach based on systematically collected corpora of naturally occurring language. This approach is a flavor of corpus linguistics, though Sampson prefers the broader term *empirical linguistics*: He considers the corpus to be the primary and essential tool for the empirical study of language.

The ambitious scope of the term *empirical linguistics* is not an accident. Introductory linguistics textbooks usually present the most fundamental distinction among schools of linguistics as that between the empiricists and rationalists. The history of twentieth-century linguistics, as usually presented, is the story of the paradigm shift from empiricism to rationalism marked by the publication of Chomsky's *Syntactic Structures* (1957). Sampson offers his empirical linguistics as an antithesis to Chomsky's generative linguistics. Indeed, in Sampson's view, a second paradigm shift has already occurred. Though "intuition-based linguistic theorizing has lingered on, in some university linguistics departments," as he puts it, empirical linguistics "began to reassert itself in the 1980s, and since about 1990 has moved into the ascendant."

One should not, however, expect from this book a sweeping, definitive exposition of the empirical linguistic paradigm. In particular, anyone seeking an introduction to recent advances in empirical computational linguistics will be disappointed. Nor is it the popularizing work that will convert the world of generative linguistics to corpus methods. It does not speak to generative linguists in their own terms, and it focuses very much on early generative linguistics: center embedding, Yngve's complexity measures (1960, 1961), Katz and Fodor's semantic marker theory (1963), Chomsky's logical structures of linguistic theory (1955 [1975]).

Rather, this book represents a particular concrete example of corpus linguistic investigation, accompanied by a critique of generative linguistics. As such, it provides some fascinating data and provocative philosophical argumentation. It consists of 10 chapters, not including the introduction. Four are empirical studies, one is mathematical, and five are philosophical.

The empirical chapters focus on depth of embedding. Chapter 2 challenges the long-standing constraint against multiple center embedding. It summarizes several variants of the constraint and presents examples from published texts that violate each variant. Unfortunately, after discarding all previous formulations, Sampson does

Downloaded from http://direct.mit.edu/coli/article-pdf/28/4/573/1797916/coli.2002.28.4.573.pdf by guest on 01 March 2024

not offer a more adequate formulation, taking instead an agnostic stance on the very existence of the constraint. Chapter 4 examines a related issue, namely, a hard limit on tree depth proposed by Yngve, and concludes that the lack of deep left recursion is not due to an Yngvean constraint on paths but rather is a consequence of the low probability of choosing left-branching expansions. Chapters 3 and 5 examine the influence of genre and social factors on depth of embedding, taking depth of embedding as a proxy for grammatical complexity. Chapter 3 looks at the effect of genre, agreeing with the common wisdom that there is a significant difference in sentence length between technical prose and fiction but concluding that it is not a consequence of a difference in overall structural complexity but is almost entirely ascribable to a difference in the number of immediate constituents in the noun phrase. Chapter 5 examines the hypothesis that social class, age, and gender have an effect on grammatical complexity finding a significant correlation only with age, not with gender or social class. Moreover, Sampson argues that there is a lifelong pattern of increasing grammatical complexity, and takes this as evidence against the existence of a "critical period" of language acquisition ending at puberty.

The mathematical chapter (chapter 7) is rather an outlier in tone and contents. Originally coauthored with William Gale, a statistician at AT&T Bell Laboratories, it gives an excellent exposition of the Good-Turing smoothing method. It presents a step-by-step recipe for computing Good-Turing discounts and an accessible but by no means trivializing account of the theory behind the method.

The remaining five chapters are more concerned with the philosophy of linguistics than with empirical investigation. Chapter 6 advocates a greater emphasis on taxonomy (particularly in the form of treebanks) in linguistics. Appeal is made to the example of biology, in which the systematizing work of Linnaeus was an essential preliminary for modern biological theory. Chapter 8 scoffs at the use of linguistic intuitions and invented examples, as opposed to corpus data, comparing a linguist using intuitions to a meteorologist who theorizes on the basis of intuitions about weather forecasting. Chapter 9 is something of an interlude, being a specific and quite detailed attack on Chomsky's *Logical Structure of Linguistic Theory*. It is perhaps best summarized by quoting a passage: "Alan Sokal and Jean Bricmont [1998] have recently documented the way that a number of twentieth-century 'intellectuals' created an air of profundity in their writings on social and psychological topics by using mathematical terminology and symbols which, to readers ignorant of mathematics, look impressive, though to professional mathematicians they are nonsensical." Chapter 10 claims that the notion of ungrammatical sentence is a Chomskyan invention, the traditional statements on the matter being in the mode "this sentence cannot be used that way," not "this sentence cannot be used." Sampson also challenges the notion of a fixed grammar, arguing that there are no invalid expansions for any category, merely a long tail of low-frequency expansions. Finally, in chapter 11, he argues that some aspects of language are beyond the limits of science. Rejecting first Katz and Fodor's formalism for lexical semantics, he goes on to reject the idea "that words have definite meanings capable of being captured in symbols of *any* formal notation" and argues that learning word meanings is more like learning to dress fashionably: There is no truth to the matter, one just tries to imitate what the more authoritatively fashionable do.

In sum, the book will appeal most to those who are interested in constraints on depth of embedding, to those interested in corpus linguistics, and to those interested in criticism of generative linguistics, particularly early generative linguistics. Anyone implementing Good-Turing smoothing will also find chapter 7 useful. It is to be recommended not as a general introduction to modern empirical linguistics, but as an exposition and example of a particularly pure strain of linguistic empiricism. To my

Downloaded from http://direct.mit.edu/coli/article-pdf/28/4/573/1797916/coli.2002.28.4.573.pdf by guest on 01 March 2024

mind, it also reveals the weaknesses of pure empiricism, especially the lack of elucidation of mechanisms giving rise to phenomena, but it certainly cannot be accused of compromising its principles.

References

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague.

Chomsky, Noam. [1955] 1975. *The Logical Structure of Linguistic Theory*. Plenum, New York.

Katz, Jerrold J. and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, 39:170–210.

Sokal, Alan and Jean Bricmont. 1998. *Intellectual Impostures: Postmodern Philosophers' Abuse of Science*. Profile. (Published in the U.S. under the title *Fashionable Nonsense*.)

Yngve, Victor H. 1960. A model and a hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.

Yngve, Victor H. 1961. The depth hypothesis. In Roman Jakobson, editor, *Structure of Language and Its Mathematical Aspects*, volume 12 of *Proceedings of Symposia in Applied Mathematics*. American Mathematical Society, Providence, RI, pages 130–138.

Steven Abney wrote this review while a principal member of the research staff at AT&T Laboratories–Research; he is now associate professor of linguistics at the University of Michigan. His interests include parsing, language learning, stochastic models, corpus methods, syntax, and semantics. Abney's address is Department of Linguistics, University of Michigan, Ann Arbor, MI 48109; e-mail: spa@vinartus.net; URL: www.vinartus.net/spa.

Downloaded from http://direct.mit.edu/coli/article-pdf/28/4/573/1797916/coli.2002.28.4.573.pdf by guest on 01 March 2024