Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English

Susan Hunston and Gill Francis
(University of Birmingham)

Reviewed by
Christopher Johnson
Soliloquy, Inc.

In this book Hunston and Francis describe an approach to lexical and grammatical description that was used to produce two remarkable Collins COBUILD reference works, *Grammar Patterns 1: Verbs* and *Grammar Patterns 2: Nouns and Adjectives* (Francis, Hunston, and Manning 1996, 1998). This approach uses large amounts of corpus data to make discoveries about lexical items and the specific phraseological and grammatical patterns in which they regularly occur. It is *corpus-driven* (as opposed to *corpus-based*) in the sense that corpus data are analyzed with minimal theoretical presuppositions about grammatical structure.

The view of language that emerges from this work is one in which no strict distinction can be made between lexicon and grammar, since lexical items must be characterized in terms of their distributions in grammatical patterns, most patterns occur with particular classes of lexical items, and many patterns are, like lexical items, specific and conventional and must be learned. The importance of this general point for computational and theoretical linguists cannot, in the opinion of the reviewer, be overestimated. Those who need to be convinced of it would probably benefit from reading this book, but should be warned: this is neither a work of computational linguistics nor of formal linguistics. Rather, it presents practical lexicographic description and is intended largely for an applied-linguistics audience (e.g., teachers of English as a second language). Computational linguists who want a source of pattern descriptions for help in designing and implementing NLP applications would be better served by the reference works cited above (but should not expect to find formal descriptions there either).

The book begins with a brief history of H&F's notion of pattern, which has its roots in the pedagogical language description of Hornby (1954) and Sinclair's (1991) work in corpus linguistics. Chapter 2 introduces patterns, mostly through examples. The patterns presented range from the very general, such as **V n** (a verb followed by a noun group), in which case they resemble basic phrase structure rules, to the quite specific—for example, *it* **V n/amount before/since** (i.e., the word *it* followed by a verb followed by a noun phrase expressing the semantic notion "amount" followed by the word *before* or *since*). Chapter 3 discusses methodological problems in identifying patterns. Chapters 4 and 5, the highlights of the book, discuss correlations between patterns and meaning, showing how clusters of semantically similar words tend to appear in the

same pattern (reminiscent of Levin's [1993] work on verb classes). Chapter 6 discusses the relation between patterns and structural descriptions that make use of grammatical function labels such as Subject and Object. Chapter 7, another highlight, discusses the way corpus patterns can be used to extend and refine traditional word classes. Chapter 8 attempts to apply the notion of pattern to the description of texts. Chapter 9 summarizes and offers some speculative claims about the theoretical significance of the pattern grammar approach.

The strongest portions of the book are the ones that present findings about specific words and patterns. Many of the observations that H&F make are insightful and truly demonstrate the value of corpus research. Chapter 5, for example, presents interesting observations about the patterns of the adjective *difficult* and the noun *difficulty*, and the differences between the words with respect to the way they allow their semantic participant roles to be syntactically expressed. This work draws on the tradition of Fillmore's (1968) case grammar, and resembles recent work by Fillmore and collaborators on frame semantics (see, for example, Fillmore and Atkins [1992]). In Chapter 7, the authors observe that the traditional distinction between count and uncount (or mass) nouns is insufficient for the description of nouns such as *injustice*, which can be used with either count or uncount properties (*crusade against poverty and injustice, acknowledging the injustices*).

As insightful as some of these observations are, there is no attempt to make sense of them in terms of an explicit theory of language. This would not be a fault if the authors were content to present their findings as purely descriptive. However, they do venture a few theoretical claims that have little foundation and in fact contradict their own descriptive practice. In their summary in Chapter 9, for example, they claim that "constituent units such as groups and clauses are challenged" by their observations about patterns (p. 259), while in fact the patterns given in the book are stated in terms of groups (phrases) and clauses. The substance of this challenge supposedly comes from Chapter 8, where patterns are applied as a descriptive tool to texts, and "linear grammar" is suggested as an alternative to constituent structure grammar. Linear grammar, however, as far as their discussion of it would suggest, makes no attempt to account precisely for linguistic structure the way a generative grammar or a natural language parser does. For example, they suggest that the pattern **V n**, which they originally defined as "verb followed by a noun group," could alternatively be defined as "verb followed by anything up to and including a noun," which is quite a different matter. In general, the authors seem to advocate a rejection of the theoretical premises of the generative linguistic enterprise (focusing on linguistic competence as a cognitive phenomenon, accounting for all and only the grammatical sentences of a language) and a return to a structuralist view of language as nothing more than a collection of utterances to be described.

There is work that addresses H&F's concerns in a more theoretically informed and formally rigorous way. Research in construction grammar, for example (see, for example, Kay and Fillmore 1999), presents explicit analyses of specific patterns that show how they relate to broad grammatical generalizations. On the computational side, there are projects (such as the COMLEX syntax—see Macleod, Grishman, and Meyers 1998) that have produced lexical resources covering large numbers of patterns like the ones observed by H&F.

While far from a definitive statement of the importance of its topic, the book does present useful examples from the perspective of two researchers with frontline experience analyzing real linguistic data. It also successfully makes the point that corpus data reveal generalizations that are difficult to find through introspective methods. Theoretical linguists will be intrigued by H&F's observations, but will likely be frustrated

by the lack of an attempt to define "pattern" rigorously and to make generalizations. Computational linguists will also long for a more formal and systematic presentation, but may nonetheless find helpful ideas in this book.

References

Fillmore, Charles J. 1968. The case for case. In Emmon Bach and Robert Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart, and Winston, New York, pages 1–88.

Fillmore, Charles J. and Beryl T. S. Atkins. 1992. Towards a frame-based lexicon: The semantics of RISK and its neighbors. In Adrienne Lehrer and Eva Feder Kittay, editors, *Frames, Fields and Contrasts*. Lawrence Erlbaum Associates, Hillsdale, NJ, pages 75–102.

Francis, Gill, Susan Hunston, and Elizabeth Manning. 1996. *Collins COBUILD Grammar Patterns 1: Verbs*. HarperCollins, London.

Francis, Gill, Susan Hunston, and Elizabeth Manning. 1998. *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. HarperCollins, London.

Hornby, A. S. 1954. *A Guide to Patterns and Usage in English*. Oxford University Press.

Kay, Paul and Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. *Language*, 75(1): 1–33.

Levin, Beth 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago and London.

Macleod, Catherine, Ralph Grishman, and Adam Meyers. 1998. COMLEX Syntax Reference Manual, Proteus Project, New York University.

Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Christopher Johnson is a member of the Core Technology team at Soliloquy, Inc., where he builds natural language understanding software for conversational agents. He received his Ph.D. in linguistics from the University of California, Berkeley, and worked as a computational lexicographer for the FrameNet project at the International Computer Science Institute. FrameNet uses corpus data to investigate the frame-semantic and syntactic properties of English verbs, adjectives and nouns. Johnson's address is Soliloquy, Inc., 251–255 Park Avenue S., New York, NY 10010; e-mail: cjohnson@soliloquy.com.