fundamental vocabulary (untouchable by the user), and "specific" vocabularies.

The traditional view of MT as a process of tree structure transformation through pattern matching is complemented by the more recent lexical approach in which the description of grammatical usage is stored in the dictionary as part of the information attached to specific words, and in which more specific rules block the more general ones (cf. Nagao 1987). Dictionary construction is therefore essential to the enterprise, and one may hope that Nagao's plea for a standardized dictionary format and international cooperation in the elaboration of lexicons will be answered by the recent dictionary and text database initiatives that have been launched to meet those needs.

As a conclusion, Chapter 8 proposes Nagao's views on the future of MT; these are summarized in the preface by Nagao's Figure 0.1, which predicts a steady improvement in the commercial systems and a sharp rise in the number of MT users during the late 1980s and early 1990s. Nagao also predicts that the improvements due to research in syntax and semantics have basically leveled off, and that further progress will come from research on intersentential components from an AI rather than purely linguistic point of view. For instance, one of the phenomena that MT cannot yet handle is discourse analysis, and both the resolution of intersentential relations and the resolution of referentials without a referent in the text require making inferences about the world. However, Nagao makes the nice point that it is very difficult for an NLP system driven by inferences to deal with new relationships set up in a text, while this is what natural language does all the time (i.e., creates novel sentences that are interpretable). High-level translation capabilities will require long-term basic research both in theoretical linguistics and in cognitive science, going beyond the limits of traditional linguistics on particular languages and toward a theory of translation.

As MT is still in its infancy, Nagao advocates a realistic assessment of its possibilities by following the "engineering practice of limitations based on assumptions about functionality." For example, the problem of voice recognition must be solved for interpreting systems, and because of efficiency considerations, the future of interactive systems probably lies in small-quantity private systems. As a final thought-provoking remark: MT is probably more useful between pairs of languages that do not have many mutual speakers, rather than the classic pairs of well-known languages.

References

Buchmann, Beat. 1987 Early History of Machine Translation. In: King, Margaret, ed. *Machine Translation Today: The State of the Art.* Edinburgh University Press, Edinburgh, U.K.

Nagao, Makoto. 1987 The Role of Structural Transformation in a Machine Translation System. In: Nirenburg, Sergei, ed. *Machine Translation: Theoretical and Methodological Issues.* Cambridge University Press, Cambridge, U.K.

Slocum, Jonathan. 1985 A Survey of Machine Translation: Its History, Current Status, and Future Prospects. *Computational Linguistics* 11(1): 1–17.

Warwick, Susan. 1987 An Overview of Post-ALPAC Developments. In: King, Margaret, ed. *Machine Translation Today: The State of the Art.* Edinburgh University Press, Edinburgh, U.K.

Dominique Estival received her Ph.D. in linguistics from the University of Pennsylvania in 1986. She gained experience in MT when developing a new version of the Weidner system at WCC. After working at Wang Labs on a large NLP system, she is now a researcher at ISSCO, where one of the projects on which she is working is the development of a prototype for an MT system based on unification. Estival's address is: ISSCO, 54 rte des Acacias, CH-1227 Geneva, Switzerland. E-mail: estival@divsun.unige.ch

---

Looking Up: An Account of the COBUILD Project in Lexical Computing

John M. Sinclair, editor
(Birmingham University)

London and Glasgow: Collins ELT, 1987, ix + 182 pp.
Paperbound, ISBN 0-00-370256-1

Reviewed by
Branimir Boguraev
IBM T.J. Watson Research Center

"COBUILD" stands for "Collins Birmingham University International Language Database," and reflects the joint nature of the work on lexical computing shared between the University of Birmingham and Collins Publishers. The COBUILD dictionary project, going back to early 1980, resulted in the publication in 1987 of the *Collins COBUILD English Language Dictionary.* There are at least three major factors that set it apart from other learners' dictionaries of English language: it is a wholly new dictionary; it reflects present-day usage of English; and, in style of presentation of entries, it represents a radical departure from existing lexicographic conventions. By its own account, "the techniques used to compile [the dictionary] are new and use advanced computer technology. For the user the kind of information is different, the quality of information is different, and the presentation of information is different" (from the introduction to the *COBUILD Dictionary*).

This difference stems from the interleaving of several basic principles in applied linguistics and dictionary compilation, and the particular ways in which these have influenced lexicographic practice in the course of preparing the dictionary. Language is a constantly changing dynamic system; consequently, no existing—and by that token already out-of-date—reference materials (including other dictionary sources) have been used in the process of compiling COBUILD. Rather, the analysis of words, from decisions concerning the makeup of the word list to the specific

content of individual entries, has been carried out entirely on the basis of studies of a large corpus of English texts, both spoken and written. As Gwyneth Fox, one of the contributors to the book, points out: lexicographers "have learned what happens when [they] sit and intuit how words are used—[they] are likely to get it wrong." This is, in a nutshell, the difference in method between "armchair lexicography" and "corpus lexicography"; and while the COBUILD dictionary remains to be judged as an artifact of the latter (see, e.g., Fillmore 1989), it is *Looking Up* that addresses the issues behind the principles, the motivations, and the realization of the corpus-based methodology of dictionary construction.

There are several main themes running through the book, recurring through most of the chapters in one way or another. The importance, to a project concerned with large-scale word studies, of corpus evidence for the way language is used within a certain time frame is clearly one of the major premises of the whole COBUILD enterprise. More than just as a side effect, this project has had to make extensive use of some specially designed computational tools, primarily for corpus maintenance and analysis, as well as for text database management. Another central notion is that of word *use,* rather than *meaning,* which is central to the dictionary design and presentation. In particular, what emerges from the corpus is a body of evidence for distinct word patterns, themselves associated with functions and uses of words. The consequences, for the whole dictionary project, can be traced in the ways this evidence has influenced decisions about how to divide a word into word senses, necessitated radically new style and conventions for explaining meaning, or constrained the choice of illustrative material (such as examples) included in dictionary entries.

The book itself roughly follows these themes, with its ten chapters falling into three major categories.

Lexical computing "Corpus Development" by Antoinette Renouf and "Computing" by Jeremy Clear address the issues of building up, from scratch, a representative corpus of English texts and the problems of storing, processing, and analyzing these in order to distill representative (and manageable) samples of word uses. The first chapter is concerned with the nature of a set of corpora (main; reserve; monitor; of spoken interaction; and so forth) required for the task of building word profiles. The second chapter presents the functionality of a set of concordancing programs (for processing and collating 7.3 million words of raw data) and of a structured database for storing the lexicographers' analysis and interpretation of the corpus data together with the evidence in support of these analyses.

Dictionary content The first task in which traditional lexicography and modern technology begin to blend is that of deciding upon the global content of the dictionary. "The Process of Compilation" by Ramesh Krishnamurthy discusses the nature of lexical data, from the perspective of what aspects of a word should ultimately be recorded in its entry, as well as the compilation and organization of such data on a large scale. In effect, this chapter sets down the "style guide" for the COBUILD dictionary. "The Analysis of Meaning" by Rosamund Moon then focuses primarily on criteria for making sense distinctions, within a word entry, on the basis of evidence found in corpus citations. In "The Nature of the Evidence," John Sinclair examines in detail the way word patterns relate to uses and meanings of individual words; even though this comes rather late in the book (Chapter 8), much of the analysis here underlies the application of the criteria set down in Moon's paper.

Dictionary presentation The remainder of the book largely concerns the ways in which the main principle of the COBUILD project has been "implemented" in the dictionary. More specifically, discussions focus on different aspects of the same question: having built a representative profile of word uses and meanings, how the relationship is best conveyed to the dictionary users. It is in the methods for expressing this, perspicuously and concisely, that the break with long-established traditions in lexicography becomes most apparent. Essential components of an entry in a learners' dictionary are its pronunciation, syntactic (grammatical) annotation, definition, and examples fields. The chapters by David Brasil ("Representing Pronunciation"), John Sinclair ("Grammar in the Dictionary"), Patrick Hanks ("Definitions and Explanations"), and Gwyneth Fox ("The Case for Examples"), thus have a common denominator: they outline notational conventions used in dictionaries to date, discuss a number of shortcomings there (some of which follow from the highly terse and compact nature of dictionary entries, and some from an inability to relate an entry to how a word is, or should be, used), and present a set of principles and/or a system for dictionary entry presentation that is intimately linked to the notion of explaining how a word influences, and is influenced by, context.

Viewed from such a perspective, the chapters in the book form a much more cohesive whole than the average edited collection. Even the last chapter, "Moving on" by Antoinette Renouf, while apparently not directly related to any part of the COBUILD dictionary project, reminds the reader that the design of that particular dictionary is only a part of a larger effort whose ultimate goal is to develop a comprehensive database of lexical data, applicable beyond the task of creating just one dictionary, and relevant to disciplines other than just lexicography. (Renouf describes the design of a series of course books, *Collins COBUILD English Course,* derived from the same text corpus.)

This book is primarily about lexicography; read in conjunction with a more or less standard text on dictionary making (for instance, Landau 1989), it demonstrates the scope and complexities of compiling dictionaries, lexicons, glossaries, and so forth. By that token alone, as well as due to the very nature of its subject matter, it should be of interest to computational linguists. However, what makes

the book more of a recommended reading is the renewed empiricism in the field, largely promoted by the very practical need to scale up natural language systems, and largely due to the realization that linguistic information about words could be derived from massive on-line text resources.

Whether these resources come in the shape of on-line dictionaries or text corpora is immaterial here. By reading *Looking Up,* one becomes acutely aware of the richness of lexical information available in (and distributed over) millions of words of text. One also understands that careful inspection of a dictionary entry (or a set of related entries) is likely to reveal considerably more in terms of lexical properties of the word (or class of words) than is apparently visible. The nature of lexical information in, and its extractability from, such text resources has been much discussed recently in the computational linguistics and computational lexicography literature; in particular, the statement that there is a wealth of implicit information available in on-line dictionaries and corpora has been made over and over again recently (see, for instance, Atkins et al. 1988; Boguraev and Briscoe 1989; Hindle 1989; Church and Hanks 1989). However, there is a world of difference between "retro-engineering," by whatever means, methods and rules for inferring lexical information from dictionary entries, and being told in advance the kinds of lexical regularities, generalizations, and properties encoded in these entries. For that reason alone, and particularly given that the COBUILD dictionary is available in machine-readable form, *Looking Up* is a book that should not be ignored by researchers interested in computational lexicography, lexical semantics, or simply the nature of word meaning.

References

Atkins, B. T., Kegl, J. and Levin, B. 1988 Anatomy of a Verb Entry. *International Journal of Lexicography* 1(2):84–126.

Boguraev, B. K., and Briscoe, E. J. 1989 *Computational Lexicography for Natural Language Processing.* Longman, London and New York.

Church, K. W. and Hanks, P. 1989 Word Association Norms, Mutual Information, and Lexicography. *Proceedings of the 27th Annual Meeting of the Associaton for Computational Linguistics,* Vancouver, B.C.

Fillmore, C. J. 1989 Two Dictionaries. *International Journal of Lexicography* 2(1):57–83.

Hindle, D. 1989 Acquiring Disambiguation Rules from Text. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics,* Vancouver, B.C.

Landau, S. I. 1989 *Dictionaries: The Art and Craft of Lexicography.* Cambridge University Press, Cambridge, U.K.

Branimir Boguraev is a Research Staff Member at the IBM T.J. Watson Research Center, where he is also in charge of the Lexical Systems Project. He holds a Ph.D. degree from the University of Cambridge, where he subsequently worked on a number of projects in natural language processing. His recent work has been in the areas of computational lexicography and lexicology. Boguraev's address is: IBM Research, P.O. Box 704, Yorktown Heights, New York 10598. E-mail: bkb@ibm.com

---

Generating Natural Language under Pragmatic Constraints

Eduard H. Hovy
(Information Sciences Institute, University of Southern California)

Hillsdale, NJ: Lawrence Erlbaum Associates, 1988, xiii + 214 pp
Hardbound, ISBN 0-8058-0248-7, $29.95
Paperbound, ISBN 0-8058-0249-5, $19.95

Reviewed by
Wolfgang Hoeppner
University of Koblenz

This book is a revised version of the author's dissertation, which was submitted to Yale University in February 1987 and published as a research report (Hovy 1987). One shouldn't be too surprised to learn from the preface that Roger C. Schank, Drew McDermott, and Bob Abelson were the honorable members of the thesis committee. Various well-known AI researchers—not exclusively from Yale—have given a hand while the thesis was on its way and are therefore mentioned in the acknowledgments. The acknowledgments, by the way, give a first example of how stylistic features affect the generation of text: Hovy switches between several styles (formality, verbosity, gratefulness, haste) while expressing his thanks to different classes of persons. The preface stresses that the book is not only useful to computational linguists, but also to theoretical linguists, especially those working in generation. Thus, the last section of every chapter deals with implementation and might be skipped by readers not interested in the computational issues.

The book describes the generation system PAULINE, which was developed and implemented by the author. The system's name is an acronym: Planning And Uttering Language In Natural Environments. (It is also the name of Hovy's sister).

Chapter 1 (11 pp.) introduces the specific research area: how do pragmatic and stylistic issues influence the generation of natural language texts? Starting with real-world descriptions of an event at Yale, the destruction of a shantytown by university authorities, it is demonstrated how different viewpoints of the respective authors affect texts. The same event is then described by PAULINE in different pragmatic adjustments. The event itself is represented in a network of about 120 elements (presumably some version of conceptual dependency), and it is claimed that the system produces over 100 different texts. The second example is a description of a fictitious primary election between Carter and Kennedy as Democratic presidential candidates. "Well, so Carter lost the primary to Kennedy by 1335 votes" is one example of a very condensed description. This terseness is beaten only by example number 12, which consists of nothing but blanks: "The program didn't find any topics that it liked and the hearer also liked,