# Book Reviews

## Computational Lexicography for Natural Language Processing

**Bran Boguraev and Ted Briscoe (eds.)**
(University of Cambridge and University of Lancaster)

London: Longman, 1989, xv + 310 pp.
(Copublished in the United States with John Wiley & Sons)
Hardbound, ISBN 0-582-02248-7 (Longman) and 0-470-21187-3 (Wiley), $49.95, £24.00

*Reviewed by*
*Geoffrey Sampson*
*University of Leeds*

As automatic natural language processing (NLP) moves out of the era of toy pilot projects and begins to grapple with real-life language in all its complexity, it needs access to quantities of information about individual lexical items. The only plausible source of such information lies in machine-readable versions of ordinary published dictionaries; although they are designed for other purposes and are far from ideal for computer use, they represent an investment of resources that the computational linguistics research community is in no position to match. This book, whose various chapters are co-authored by a total of sixteen contributors drawn from research groups at Cambridge, Amsterdam, and New Mexico State University, is about the issues that have arisen in exploiting one dictionary, the *Longman Dictionary of Contemporary English* (*LDOCE*), for NLP purposes.

Topics covered include: the translation of what was originally a typesetting tape into a format enabling the computer to locate efficiently the various categories of information specified for a given word; the relationship between grammatical information as listed for entries in the dictionary and the categories required by the theoretical linguist; the use of pronunciation information in speech-recognition systems; and various tentative experiments in deriving computer-usable semantic information from the definition portions of dictionary entries. A series of appendices gives useful factual information about the dictionary and its vocabulary. Since all the work discussed relates to *LDOCE*, it concerns English exclusively; at one point (p. 8) the editors offer a French example that is truly hair-raising in spelling and pronunciation, but this is a rare blemish in a book whose standard of editing and production is high.

The research reported has a solid and professional quality. These are no dilettantes musing on how in principle one might aim to achieve a given task, but people who have set about getting the job done, for the whole language, and tell us here what problems they encountered and how far they were successful. Most of the problems will be common to any published dictionary, and the introductory chapter and the bibliography give very full coverage of dictionary-based NLP research in general; for a research group aiming to work with a different English dictionary, or a dictionary of another language, this book would be an excellent place to start. Although much of the emphasis is rightly on practical computing considerations, there are also many novel and valuable theoretical insights. David Carter, discussing speech-recognition systems that use partial phonetic information to produce classes of candidate words, points out that standard ways of measuring the systems' performance are grossly misleading, both because they ignore relative frequencies of words in a class and because measurement ought to be logarithmic rather than linear—to whittle a vocabulary of 10,000 words down to a class of 100 candidates is to do half, not 99%, of the task of identifying the stimulus. Theoretical linguists have claimed that various aspects of grammatical subcategorization of lexical items are predictable from meaning and thematic structure, as one might well expect (how can a person master a language unless such matters are governed by general rules?), but Boguraev and Briscoe exploit that fact that *LDOCE* codes individual verb-senses for susceptibility to dative alternation (*Sally slid the drink to Susan* → *Sally slid Susan the drink*) in order to submit this idea to detailed testing, and they find that none of the proposed rules holds; whether dative alternation is permissible seems to be an aribtrary fact about individual senses of individual words (a point that underlines the crucial importance of full-scale dictionaries in NLP).

Rather than discussing each chapter of this book in detail, in the rest of my review I shall take up two issues that recur in many chapters and on which I am inclined to question the position put forward by the contributors.

The first has to do with how far an ordinary published dictionary comparable to *LDOCE* (which has 55,000 entries) provides adequate coverage of text, in view of the open-ended quality of any natural language and the fact that published dictionaries do not try to cover proper names. Many contributors to the book reviewed are quite pessimistic, suggesting that while published dictionaries are the best resources we have, they fall very far short of adequate coverage of the language. The writers repeatedly refer to Walker and Amsler (1986), who studied the extent of correspondence between a large sample (over 8 million word-tokens) of material from the *New York Times* news service, and the (approximately 70,000) entries of *Webster's Seventh New Collegiate Dictionary*. Walker and

Amsler say that as many as 64% of the word-types in their corpus do not occur in the dictionary.

This surprised me, since it seemed to contradict the findings of a smaller-scale piece of research that I carried out (Sampson 1989) before learning of Walker and Amsler's work. I ran the computer-usable version prepared by Roger Mitton (1986) of the third edition of the *Oxford Advanced Learner's Dictionary* (*OALD3*) against a 45,622 word-token subset of the LOB Corpus, in order to analyze in detail the nature of the gaps in the dictionary (like Amsler and Walker, I eliminated punctuation, "words" consisting of digits, etc.). Only 1,477 word-tokens, or 3.24% of my sample, were missing from the dictionary. Although I was counting tokens while Walker and Amsler counted types, in itself this can hardly explain more than a fraction of the difference between our results. Zipf's Law, taken in conjunction with the rank/frequency information about the LOB Corpus in Hofland and Johansson (1982), suggests that 44,145 tokens found in the dictionary might represent on the order of 18,000 types, and the 1,477 missing tokens represent 1,176 types, implying that in my sample the non-dictionary word types would be in the region of 6%. Walker and Amsler's dictionary lacked inflected forms and also some high-frequency proper names (though its total number of entries is nevertheless slightly lower than that of the Webster dictionary); furthermore, the news service corpus is likely to have a higher density of proper names than the LOB Corpus. But these factors do not account for the difference between Walker and Amsler's and my figures; they state that inflected forms and proper names comprise only half of their nondictionary word types, so that about one-third of all word types in their sample is missing from their dictionary without being names or inflected forms. (Because of the large scale of their experiment, Walker and Amsler are not in a position to give detailed analyses of this one-third.)

I can only suppose that the discrepancy between Walker and Amsler's figure of one-third and my 6% stems from the fact that relative size of sample matters when counting missing types, rather than missing tokens. Even if Walker and Amsler had no greater a proportion of nondictionary tokens than I, in their sample that would be a quarter of a million tokens, many of which would be unique types, while for either sample the bulk of tokens that are in the dictionary will be concentrated on the same few very common types. If Walker and Amsler had counted tokens rather than types, then (and had used a dictionary that listed inflected forms), it is not clear that they would have reached a figure much higher than my 3.24%. This puts a different gloss on the question of dictionary adequacy.

The second issue I should like to take up is the claim, made by several contributors to this book, that *LDOCE* is "uniquely suitable for computational lexicography" (p. 2). There has been an idea in the air for some time now that *LDOCE* is not just one lexical resource among others but, for computer applications, has a clear superiority over all alternatives. The point has not been easy to assess, since it relates in part to material found only in the electronic version of *LDOCE*, and (unlike Oxford, who have made *OALD3* relatively freely available to computational researchers), Longman have strictly guarded their electronic copyright, allowing the dictionary to be used only by a few groups whose work is centered on it and who as a natural consequence have tended to champion it against its rivals. My own loyalties lie rather with *OALD*, edited by my Leeds/CCALAS colleague Anthony Cowie; this book offers an opportunity to try to establish how much of the claimed superiority of *LDOCE* is real and how much hype. (I shall compare *LDOCE* only with *OALD*; the book under review itself treats *OALD* as "the competition," and I have little experience with other dictionaries such as Webster's or the Collins/University of Birmingham *COBUILD* dictionary.)

Relevant points made by contributors fall under four headings: file format; "controlled vocabulary"; subject and "box" codes; and grammatical classification schemes.

So far as file format is concerned, Eric Akkerman claims (p. 66) that "the computer-tape version of *LDOCE* turned out to be much more structured than that of *OALD*, which was basically a typesetting tape." This is certainly fair comment with respect to the tapes originally produced by the two publishers. For some time now, though, it has been of historical interest only, since more than one research group have produced parsed versions of the *OALD3* tape. Furthermore, in this and other respects the critique of *OALD* in this book has been overtaken by the publication, also in 1989, of the fourth edition of *OALD*, which was designed from the start to be computer-tractable and has a format that I suspect is superior to that of the *LDOCE* tape (though, certainly, by no means as sophisticated as that of the resource that some of these researchers have created using *LDOCE* as their raw material).

"Controlled vocabulary" refers to the fact that the language of *LDOCE* definitions is restricted to a specified set of some 2,200 words, each of which is supposed to be used only in its central sense(s). This might make it easier to deduce formalized representations of word meanings from the definitions, and several of the contributors who work on computational semantics express enthusiasm for this feature of *LDOCE*. The editors are much more cautious, however (see pp. 16, 34); and research by Jansen *et al.* (1987) makes it questionable how far one can accurately describe the *LDOCE* definition vocabulary as "controlled" at all. Furthermore, an explicit controlled-vocabulary policy is presumably an advantage only if, without it, lexicographers tend to range more widely in phrasing their definitions. To test this, I looked at the examples of *LDOCE* definitions quoted by Hiyan Alshawi (p. 157ff), who has developed a system that attempts to assign word senses to general semantic categories by locating the grammatical head word of the definition, commonly a superordinate of the definiendum. In many cases the same head word occured in both *OALD4* and *LDOCE* definitions; and in every case the *OALD4* head word seemed at least as intuitively

suitable as that in *LDOCE* as a core-vocabulary superordinate term—in the case of *club* (verb), one might think that *OALD4 hit* is more straightforward and unambiguous than *LDOCE beat*. *OALD4* has no definition for *bring out* corresponding to *LDOCE's* "to introduce (usu. a young lady) into the social life of a great city."

Probably more important are the "subject" and "box" codes, which are included (in the subentries for individual word senses) only in the electronic *LDOCE*, not in the published version; this book seems to be the only public source of information about them. Subject codes are four-letter sequences that represent the topic and/or geographic domain in which a word or word sense is used, e.g. NAZV "nautical," GAQA "games/Argentina." The ten-character box codes express semantic selection restrictions; an example from page 14 is --L-X----S for *sandwich* "to put tightly in between," where *X* in place 5 means preference for abstract or human subject, and *S* in place 10 means preference for solid object. Unfortunately, very little detail is given about the box code system; thus, although place 3 is said to be associated with "level of attitude," we are not told what the *L* in the *sandwich* box code means, and in general there is no statement of the range of categories expressed by these codes. Subject and box codes both look as though they might be genuinely significant NLP resources having no equivalent in other dictionaries; the subject codes might be usable for disambiguating words in context, the box codes could serve that purpose and also, conceivably, might make a contribution in grammatical parsing. But the book does not tell us enough to permit their value to be assessed, and the editors note that "there are problems concerning the accuracy and completeness of this type of information in *LDOCE*. . . . None of the work reported in this book makes significant use of these codes."

The singlemost important reason for the book's claims about the superiority of *LDOCE* is its system of grammatical classification; Chapter 3, by Eric Akkerman, is an extended comparison between the classification schemes of *LDOCE* and *OALD3*, very much to the latter's disfavor. *LDOCE* uses a system of two- or three-part codes in which, broadly, a capital letter indicates part of speech and a digit indicates valency; in some cases a lowercase letter is added to indicate a finer subdivision. Thus the use of *know* as in "I know (that) he'll come" is coded T5a, in which *T* means "transitive verb with one object," 5 means "followed by a *that*-clause," and *a* means "*that* can be omitted." *OALD3* uses intuitive abbreviations (*n*, *vt*, *prep*, etc.) for the sort of information coded by capital letters in *LDOCE;* verb valencies are indicated by specifying one or more of the 51 "verb patterns" of a system worked out by A. S. Hornby, the first editor of *OALD*. The use of *know* just quoted is VP9, i.e., "S + *vt* + *that*-clause." One respect in which the *LDOCE* system is more flexible is that its valency digits can combine with letters for any part of speech to which they apply, not just with verb letters; thus 5 combines also with F (attributive adjective or adverb) to give a code F5a for *sure* as in "He was sure (that) she knew." Only a few combinations of nonverb letter with number are meaningful, but those that are give information that is not coded in *OALD* at all.

It is also fair to say, as Akkerman does, that the Hornby verb-pattern system is in some cases linguistically naive, failing to distinguish cases in which items occur together through a grammatical construction from cases of accidental juxtaposition. He quotes VP14A, "S + *vi* + *to*-infinitive," used in *OALD3* both for *come* in "How did you come to know her?", where the *to*-clause is in construction with *come*, and for *stop* in "We stopped to have a rest," where the *to*-clause is an immediate constituent of the sentence. For reasons like this, the Hornby system has been abandoned in *OALD4*, which uses a more rational method of classifying verbs (it is also more advanced than the *OALD3* system in other areas of grammar; for instance, it codes nine grammatical types of noun rather than just dividing nouns into countable and uncountable as in *OALD3*). Working through Akkerman's detailed comparison of the two dictionaries, I find that in most of the examples he quotes to illustrate the superiority of *LDOCE* over *OALD3*, the *OALD4* coding of the relevant word senses escapes his criticism.

On the other hand, Akkerman also appears to me to be less than fair even to *OALD3*. He often seems to begin the axiom that the *LDOCE* grammatical distinctions are ideal, so that any deviation in *OALD*, whether by conflating classes that *LDOCE* distinguishes or by drawing distinctions that *LDOCE* does not make, must be bad. He complains (p. 75) that *OALD3* has no code corresponding to the *a* of *LDOCE* T5a, showing that *that* can be left out from a finite clause following the verb; but this code is useful only if verbs differ with respect to omissibility of *that*, and omission of *that* is normally governed by considerations other than verb identity. (I thought *ascertain* might be a verb not tolerating omission of *that;* but *LDOCE* codes it as T5a.) Conversely, Akkerman says that there is no grammatical motivation for the Hornby-code distinction between verbs of physical perception and other verbs; but in (British) English the former verbs are grammatically distinctive in requiring *can* in the present tense ("I can see the car" but not *"I see the car"), though I am not sure whether this explains the feature of the Hornby system to which Akkerman refers. At one point (p. 74), Akkerman even criticizes the *OALD3* system as inferior to that of *LDOCE* by reference to a point on which they seem to agree. Akkerman says that it is illogical for *OALD3* to use a single Hornby code VP19C to cover the two uses of a verb like *understand* as in "I can't understand him behaving so foolishly" and "I can't understand his behaving so foolishly," arguing that the syntax of the two examples is very different. Again it is not clear that verbs differ in their propensity to occur in one of these patterns rather than in the other; but, more to the point, I can find only a single code, T1, applicable to either pattern in the *LDOCE* entry for *understand*.

In sum, I am not convinced that researchers without access to the Longman dictionary are doomed to second-class citizenship in the emerging world of natural language

processing. *LDOCE* has some attractive special features, but so do other dictionaries. There can be little doubt, on the other hand, about the importance and value of the kind of research reported in this book.

## References


Hofland, Knut and Johansson, Stig. 1982 *Word Frequencies in British and American English.* Norwegian Computing Centre for the Humanities, Bergen.

Jansen, J.; Mergeai, J.P.; and Vanandroye, J. 1987 Controlling *LDOCE*'s Controlled Vocabulary. In: Cowie, A.P., ed., *The Dictionary and the Language Learner* (Lexicographica, series maior, 17). Niemeyer, Tübingen, 78–94.

Mitton, Roger. 1986 A partial dictionary of English in Computer-Usable Form. *Literary and Linguistic Computing* 1:214–215.

Sampson, G.R. 1989 How Fully Does a Machine-Usable Dictionary Cover English Text? *Literary and Linguistic Computing.* 4:29–35.

Walker, D.E. and Amsler, R.A. 1986 The Use of Machine-Readable Dictionaries in Sublanguage Analysis. In: Grishman, Ralph and Kittredge, Richard, eds., *Analyzing Language in Restricted Domains: Sublanguage Description and Processing.* Lawrence Erlbaum, Hillsdale, NJ: 69–83.



*Geoffrey Sampson* is Professor of Linguistics and Director of the Centre for Computer Analysis of Language and Speech at the University of Leeds, Britain's largest unitary university. His computational linguistics research is corpus-based, and includes development of a system of robust parsing by stochastic optimization (Project APRIL). With Roger Garside and Geoffrey Leech he coedited *The Computational Analysis of English* (Longman, 1987). Sampson's address is: Department of Linguistics and Phonetics, University of Leeds, Leeds LS2 9JT, U.K.


---

## Natural Language Processing in LISP: An Introduction to Computational Linguistics

## Natural Language Processing in POP-11: An Introduction to Computational Linguistics

## Natural Language Processing in PROLOG: An Introduction to Computational Linguistics


**Gerald Gazdar and Chris Mellish**
(University of Sussex and University of Edinburgh, respectively)



Workingham, England: Addison-Wesley, 1989
Lisp volume: xv + 524 pp.
Hardbound, ISBN 0-201-17825-7, £17.95
Pop-11 volume: xv + 524 pp.
Hardbound, ISBN 0-201-17448-0, £17.95
Prolog volume: xv + 504 pp.
Hardbound, ISBN 0-201-18053-7, £17.95



*Reviewed by*
*Kwee TjoeLiong*
*University of Amsterdam*


This is a very interesting and intriguing array of textbooks to read, to compare, and to review. It also has been a rather hard job for me to do so. The last paragraphs try to explain why. First of all, however, an objective and factual summary of the contents and form of Gazdar and Mellish's *NLP in X: An Introduction to CL,* where *X* is instantiated to one of {*PROLOG, POP-11, LISP*} (reviewer's shorthand).

Quotations can be helpful as the shortest way to give you a rapid impression. From the letter that the book review editor sent me is this encouraging line: "They are really three separate versions of the same book, so there's not nearly as much reading as there first appears." Therefore, one of the titles is treated here as prototypical (to wit, the Prolog volume). Whenever they differ, the other two are, subjectively, considered as derivative.

I am going to quote amply from the authors' Preface, since it is a characterization of the book in their own words. It is neatly split into sections, and I distinguish three aspects: 'What,' 'What Exactly,' and 'In What Way,' each aspect being handled in a pair of consecutive sections of the preface.

**What:** From the first two sections, *Audience* and *Coverage:*

> This book is aimed at computer scientists and linguists at undergraduate, postgraduate or faculty level, who have taken, or are concurrently taking, a programming course in *X*. . . . The book is specifically intended to teach NLP and computational linguistics: it does not attempt to teach programming or computer science to linguists, or to provide more than an implicit introduction to linguistics for computer scientists. . .
>
> The major focus of this book, as of the field to which it provides an introduction, is on the processing of the orthographic forms of natural language utterances and text. [No issues in speech, because those are] topics that deserve books to themselves, books that we would not be competent to write. Most of the book deals with the parsing and understanding of natural language, much less on the production of it. This bias reflects the present shape of the field, and of the state of knowledge. . .
>
> The book is formally oriented and technical in character, and organized, for the most part, around formal techniques. The perspective adopted is that of computer science, not cognitive science. . . . We concentrate on areas that are beginning to be well understood, and for which standard techniques . . . have begun to emerge. . . . [Hence,] a good deal more time on syntactic processing than on semantic or pragmatic processing. . . . Discussion of developments at the leading edge of NLP research, on such topics as parallel parsing algorithms, the new style categorial grammars, connectionist approaches or the emerging implementations of situation semantics and discourse representation theory are excluded altogether or relegated to the further reading sections. . . . A less readily excusable omission is any consideration of the role of probabilistic techniques in NLP. [But . . . ]