Electric Words: Dictionaries, Computers, and Meanings

Yorick A. Wilks, Brian M. Slator, and Louise M. Guthrie
(University of Sheffield, Northwestern University, and Lockheed Martin)

Cambridge, MA: The MIT Press
(ACL–MIT Press Series in Natural
Language Processing, edited by
Aravind K. Joshi, Karen Sparck Jones,
and Mark Y. Liberman), 1996,
ix+289 pp; hardbound, ISBN
0-262-23182-4, $32.50

Reviewed by
Archibald Michiels
University of Liège

A funny title—I surmise that it will often be misquoted as *Electronic Words*. Is there a hidden citation behind it? I haven't been able to trace it.[1] *Electric Words* (henceforth *EW*, also used to refer jointly to the three authors) is a report on work done to, and with, machine-readable dictionaries, in particular LDOCE, the *Longman Dictionary of Contemporary English* (1978 edition). **To:** Machine-readable dictionaries are often nothing else than typesetters' tapes, a far cry from lexical data bases. Somewhere in the middle is *EW*'s concept of machine-tractable dictionaries, where the information is formalized to a certain extent, the extent depending on the nature of the information itself, ranging from the easily formalizable (because simply listable: parts of speech, subcategorization codes) to unformalizable (genuine citations, i.e., unrestricted natural language). **With:** Machine-tractable dictionaries have been—and are still being—used in a wide range of NLP applications, such as full-text information retrieval and machine-assisted translation.

EW offers a world-wide survey, even if the stress (somewhat inevitably) falls on the work carried out at the Computing Research Lab of the New Mexico State University, to which the three authors were attached at the time of writing (1993–94). The survey is carried out on historical principles, an orientation that is bound to be gratifying to the early workers in the field, who for once get the credit they are due. The historical perspective also has a sobering effect, showing that the key issues (and even sometimes techniques) can be traced back to prehistoric times (the sixties—the work carried out by Karen Sparck Jones is an apposite example).

I do not think there is much point in giving a summary of what is already a

1 In accordance with the usage note of the *Collins English Dictionary s.v. electronic*: "*Electronic* is used to refer to equipment, such as television sets, computers, etc., in which the current is controlled by transistors, valves, and similar components and also to the components themselves. *Electrical* is used in a more general sense, often to refer to the use of electricity as a whole as opposed to other forms of energy: *electrical engineering; an electrical appliance. Electric,* in many cases used interchangeably with *electrical*, is often restricted to the description of particular devices or to concepts relating to the flow of current: *electric fire; electric charge.*"

Graeme Hirst has suggested to me that "in the late 1980s, the use of *electric* in place of *electronic* was a vogue in English, the connotation being to make a high-tech object seem more familiar and friendly." He has also suggested that the allusion might be to the title of Louis Rossetto's now-deceased language-technology journal *Electric Word* (the name of which survives as that of the news-briefs section of his current journal, *Wired*).

summary. *EW*'s purpose is not to report on new research,[2] but rather to ground the research reported on in a philosophical perspective. The main body of the book can be divided into three parts:

- A historical overview of theories of meaning, carried on in a spirit of philosophical inquiry (Chapters 2–4). This section is often polemical in tone, and Yorick Wilks's hand is easy to recognize (the abrasive humor alone would betray the author). The main tenet of the position argued for is that meaning is symbolical ("meaning is other words or symbols," p. 15). From there it follows that primitives are words (of a given natural language, be it English or Swahili or whatever).
- A look at dictionaries and thesauri as products of the lexicographical tradition, which relies on a number of central hypotheses, the most important of which is the divisibility of a word's semantic space into word senses (Chapter 5).
- The transformation of machine-readable dictionaries into machine-tractable dictionaries and an overview of their uses in NLP (Chapters 6–14). This part includes an interesting discussion of "empty" heads in dictionary definitions and of the *genus-differentiae* definition pattern in general, as well as of genus disambiguation techniques (Chapter 10).

I think it would have been useful to give the reader a pointer to work that pursues similar aims without making use of dictionaries at all (pure corpus-based research). I suggest Grefenstette (1994)[3] as a good starting point. The reference would have fitted nicely in the discussion of the Pathfinder Networks (p. 116 and again p. 128). Similarly, as a complement to the discussion of the COBUILD dictionary and the corpus it is based on (p. 101), it is now possible to refer the reader to a very useful CD-ROM, COBUILD COLLOCATIONS.

A survey such as *EW* cannot afford to go into details, so it would be wrong to insist on an in-depth treatment of the issues the authors raise. But they can be blamed for overusing the phrases *in depth* and *in detail*, when all they provide is a cursory treatment. For instance, on page 125, they claim that the LDOCE codes are described "in detail" in Chapter 7, where they are assigned a single paragraph (p. 99). On page 125 the description is really too meager to be of much use. Besides, the authors are liable to mislead the reader when they write "It is usual for verb entries to have more than one of these codes." The important thing to note is that the codes are assigned to word senses, not to whole entries. Without this knowledge it is difficult to understand the use (discussed in *EW*) made by Boguraev and Briscoe (1987) of LDOCE grammatical code assignments to provide a classification of verbs into such classes as subject-raising verbs, object-equi verbs, etc., on the basis of "suggestions" to be found in my doctoral dissertation (and further discussed in Michiels 1995).

In general I am very much in agreement with the positions argued for by the three authors. One area of disagreement is the assessment of the value of LDOCE for

2 Which doesn't mean, of course, that the reader won't find anything new for him or her. I have profited from the very interesting exposition of the technique of **simulated annealing** (pp. 202–6). I think a similar treatment ought to have been given to the notion of **mutual information** (which is mentioned rather than discussed, p. 185), insofar as it has proved crucial for the study of collocations.

3 *EW*'s bibliography has only one entry for 1994, so it is probably safer to think of 1993 as bibliographical *terminus ad quem*.

some of the research they report on. I think they tend to minimize the problems raised by the use of a controlled defining vocabulary. In LDOCE lexical "simplicity" is often bought at the price of syntactic convolution and unnaturalness. I very much doubt that sense-tagging the defining words in LDOCE definitions always makes sense or even is always possible. The authors claim (p. 201) that LDOCE's small defining vocabulary makes it a useful corpus for obtaining co-occurrence data. But being compelled to use a controlled vocabulary has often led the LDOCE lexicographers to write definitions from which it would be very hard to see which word or word sense is being defined (as in the eighth definition of *keep*: "to have for some time or for more time"). It is not clear that the data that one can extract from such definitions are really relevant for co-occurrence studies, if by these are meant studies of collocational properties. Also, my experience with the LDOCE tape has taught me that LDOCE semantic features are often assigned rather erratically, and that the LDOCE semantic hierarchy itself, with its inclusion of funny *or*-features, is not carefully thought out. The authors are more clearly aware of the danger of using LDOCE examples, which are not genuine, but coined ("concocted with non-native English speakers in mind," p. 196).

The survey presented in *EW* cannot be expected to extend up to the year of publication of *EW* itself, 1996. But I think the reader is entitled to know the upper time limit, and the introduction ought to have mentioned it.[4] From the bibliography it would seem that 1993 is the real border, as already mentioned in footnote 3. It is not surprising, then, that the chapter that already feels the most dated should be the last, entitled "The present." In this chapter the authors report on E.C.-funded R&D projects (p. 245). Besides uncritically repeating the proffered aim ("can be considered as a follow-up to the Eurotra MT project in an effort to make some use of the lexicons constructed as part of that enormous enterprise"), they draw up a catalogue of things to be done (of very little interest unless one is writing a history of European NLP policies) rather than report on work actually carried out. No fewer than four times the reader is faced with a curious use of *is to* in the preterite (e.g., "ET-10 63 was to enhance the Eurotra system"), which one has the choice of interpreting as either some kind of epistolary past—as when Cicero ends a letter to Atticus with the word *Valebam*, meaning *I was in good health (when I wrote this)*—or as an uncanny report on the project's suspected failure (*was to, but didn't*). The *Cambridge International Dictionary of English* (CIDE 1995), which the authors do not refer to by name but describe as "the new dictionary from Cambridge," i.e., forthcoming, has been sitting on my desk for quite a while (it came out in the second half of 1995).

A survey of work carried out on machine-readable dictionaries in the years 1993–96 would probably lay greater emphasis on bilingual dictionaries (cf. the E.C.-funded DECIDE and COMPASS projects). The judgment passed on such dictionaries in *EW*—despite the hedges—sounds a bit harsh: "A printed bilingual dictionary is normally little more than lists of pairs of equivalent strings from different languages, usually with some examples of usage mixed in" (p. 208). Good bilingual dictionaries, such as the Collins bilinguals and *Oxford-Hachette* (OH 1994), display impressive metalinguistic apparatus (collocate lists, field labels, grammatical environments, etc.).

On the whole, I find *EW* a fair and stimulating survey, and I recommend reading it from cover to cover, as I did for the purpose of writing this review. Anybody who does that will gain an understanding of the philosophical underpinnings of the lexicographical enterprise and the computational use of its products.

4 Even more so since there is at least one "at the time of writing" reference in the body of the text (p. 109).

I must now turn to my main criticism of *EW*, which concerns the lack of care with which it was written and put together. The authors were wrong to skip lightly over—not to say skip *tout court*—the necessary final stages of book production, which are always of crucial importance, and even more so when more than one writer is involved.

The contributions have not been carefully integrated into a final product, which should read like a seamless whole. On the contrary, one all too readily feels the lack of an overseer. One example among several is the discussion of WordNet. It is distributed over several sections, which would be acceptable if one felt from that distribution of the information that that is the best way of building a complex picture. But what we have is first a general presentation and discussion on pages 126–7, followed on page 145 by a much shorter introduction, where the authors do refer back to the earlier discussion, but at the same time give the impression that they do not know (or have totally forgotten) what is written there. Another striking example of this lack of integration is provided by the discussion of the ACQUILEX project. On pages 209–11, we get a discussion of the creation of the ACQUILEX Lexical Knowledge Base and of the so-called *grinding* rules (a perceptive and critical assessment), without any reference to anywhere else in *EW*. On page 245 ACQUILEX I and II are briefly mentioned, with a "were discussed earlier in detail" warning, but there is no reference to pages 209–11. On pages 246–7 we get a new subtitle, "ACQUILEX", and a very general overview of the project's aims, which is what one would have expected to start with (and again, no pointer to anywhere else in *EW*).

EW also exhibits problems with "cut-and-paste" passages, probably from earlier contributions by the authors. A book ought to be written on the negative impact of word-processing systems on writing skills! Instead of rewriting (which is often accompanied by some amount of rethinking), we all tend to import and inject into our texts bits that were written for other purposes. They may be adequate from a strictly "content" point of view, but they often carry over telltale signs of the context they were embedded in. A clear example in *EW* is the final paragraph of Chapter 10 (p. 181), which is not a conclusion at all but an abstract. The chapter is devoted to "Genus hierarchies and networks," which of course includes a full discussion of the derivation of semantic networks from LDOCE. The paragraph in question reads as follows:

> Automatic techniques for selecting the genus term in an LDOCE definition and disambiguating it relative to the senses of LDOCE have proved very effective. A hierarchy of 39,000 nouns and phrases defined in LDOCE was constructed using these results. Analysis of this hierarchy shows that it is relatively shallow (the median depth of a node is two levels down), but this is mostly a consequence of restricting the defining vocabulary to a small set.

EW is marred by a number of stylistic infelicities. In the best of cases, they are only annoying (like typos, to be discussed below); at worst, they prevent understanding. Of the former type are the following:

- "[Dictionaries] wait almost all their useful lives on shelves waiting to be all too briefly consulted" (p. 75);
- "As for the corpus being used to decide how to partition a word into senses, it should be noted that although the corpus helps the lexicographer to partition a word into senses, . . . " (p. 102).

Of the latter type:

- "Other researchers believe that dictionaries may indeed contain sufficient knowledge, although that knowledge may be implicit, because that knowledge can be made explicit . . ." (p. 142);
- "the gathering aspect is called the Cambridge Language Survey" (p. 244);
- "Various percentages of the system's semantic lexicon were reduced" (p. 252).

It may seem a petty cavil to bring up the question of typos,[5] but *EW* exceeds the tolerance threshold for a published book, let alone an expensive hardback. What is acceptable for pre-publication conference proceedings is not necessarily acceptable in a finished product. Particularly exceptionable are the typos that a simple-minded spelling checker would normally catch: *be be* (p. 3), *nineteeth* (p. 65), *concerened* (p. 86), *examing* (p. 107), *accomodated* (p. 125), *aquisition* (p. 131), *prounounciation* (p. 139), *co-ocurrences* (p. 152), *co-ocurrence* (p. 186), *exisiting* (p. 243), *sematics* (p. 274). My favorite is *Bolf-Berunek and Newman* (p. 270): has anybody actually proofread that?—it sounds as if it came straight out of a (highbrow?) sitcom. In at least one case the reader will undoubtedly want to know whether the typo is in the corpus or in *EW*'s printing of the corpus (*edage* for *edge* in the first corpus line of Figure 11.1, p. 187).

What's worse, some typos might leave the reader somewhat puzzled. Let the reader make up his or her own errata list with the following: *word* (p. 40, l. 4 → *wood*), *and* (p. 103, l. 25 → *were*), *I* (p. 140, l. 8 → *I0*), *described* (p. 234, l. 1 → *derived*). And I am still puzzled by the following (should *later* read as *larger*?): "Grammatical information is secondary, but also useful, particularly for the predictions of later grammatical constructions that are made possible by the LDOCE comprehensive grammar for English" (p. 235).

The bibliography would have been clearer if a separate section had been devoted to dictionaries and corpora. COBULID is mentioned on page 5 and again on page 64, but is not a bibliographical entry. We have to wait until page 98 to have a reference to Sinclair 1987a, which has its entry in the reference list (although the text associated with the entry has an intrusive *In* as if it were a paper in a book). The reference list is defaced by a number of typos, some affecting authors' names (*Schuetze* instead of *Schutze*, *Morgeai* instead of *Mergeai*). There are also missing entries: *Jansen et al. 1985* (referred to on page 209), *Krovetz 1992* (p. 98).

The index seems reliable enough, but I have been unable to find the rationale followed by the authors for the inclusion of people's names. At some point, I thought I had found out: proper names are included in the index when they are mentioned in the body of the text without being included in a bibliographical reference (e.g., *Grace Murray Hopper*, p. 28). But pursuing that hypothesis further on in the book proved it wrong (*Guzman* and *Brady* on page 107, *Marcus* on page 108 are neither in the bibliography nor in the index).

In conclusion: *EW* is a valuable survey, but was much too hastily written and put together.

5 Although I feel entitled to protest against the misspelling of my own name (*Michiel* on page 86), even if I am in good company (*Borguraev*, on page 143, twice!).

References

Dictionaries and corpora

CIDE = Procter, Paul, editor-in-chief. 1995. *Cambridge International Dictionary of English*, Cambridge University Press.

COBUILD = Sinclair, John, editor-in-chief. 1987. *Collins Cobuild English Language Dictionary*, first edition, Collins, London and Glasgow.

COBUILD COLLOCATIONS = *Cobuild English Collocations on CD-ROM*, 1995, HarperCollins. *Collins English Dictionary* = *Collins English Dictionary*, 1994, Third edition, HarperCollins.

LDOCE = Procter, Paul, editor-in-chief. 1978. *Longman Dictionary of Contemporary English*, Longman Group Ltd, Harlow, England.

OH = Corréard, Marie-Hélène and Valerie Grundy, editors. 1994. *The Oxford-Hachette French Dictionary, French-English, English-French*, Hachette and Oxford University Press.

Other references

Boguraev, Branimir K. and Ted Briscoe. 1987. Large lexicons for natural language processing: Utilizing the grammar coding system of LDOCE. *Computational Linguistics*, 13(3–4):203–218, September–December.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Press, Boston.

Michiels, Archibald. 1995. Retrieving verb classes from LDOCE. In Karl Sornig et al., editors, *Linguistics with a Human Face. Festschrift für Norman Denison zum 70. Geburtstag*. Grazer Linguistische Monographien, 10, Graz, pages 223–233.

Archibald Michiels wrote his Ph.D. thesis (1982) on LDOCE and its potential exploitation as an NLP resource. He is currently Professor of Computational Linguistics at the University of Liège. Michiels's address is: English Dept., University of Liège, 3, Place Cockerill, B-4000 Liège, Belgium; e-mail: amich@vm1.ulg.ac.be