MACHINE-READABLE COMPONENTS IN A VARIETY OF INFORMATION-SYSTEM APPLICATIONS

Howard R. Webber
Reference Publishing Division
Houghton-Mifflin Company
2 Park Street
Boston. MA 02108

Components of the machine-readable dictionary can be applied in a number of information systems. The most direct applications of the kind are in wordprocessing or in "writing-support" systems built on a wordprocessing base. However, because a central function of any dictionary is in fact data verification, there are other proposed applications in communications and data storage and retrieval systems. Moreover, the complete interrelational electronic dictionary is in some sense the model of the language; and there are, accordingly, additional implications for language-based information search and retrieval.

In regard to wordprocessing, the electronic lexicon can serve as the base for spelling verification (in which the computer detects many spelling or typographical errors) and spelling correction (in which the computer offers corrections to the errors it has identified). Because it is possible to develop algorithms that permit the computer to calculate the chances that the single best alternative it offers is actually correct, this substitution can in many cases be made automatically. It is at this point in the development of such systems wise to flag such automatic corrections for inspection by the operator.

At the present time, these processes generally depend upon the application of strict frequency measures, which permit the lexicon to be reduced to small-machine proportions and thereby reduce the possibility of a false hit--the passing of a misspelled common word that happens to coincide in orthography with a legitimate but rare word. As our ability to draw cognitive information from text increases, and as available memory increases, then such limits can be abandoned.

Truncation of the lexicon for other specific applications can be considered. It is possible, for example, to shape the lexicon to reflect a children's vocabulary and thereby to develop spelling correction and other writing aids for the early educational years on a very small machine base. It is also possible to shape the lexicon to the needs of the educated adult user, for whom information about common words is unnecessary, and thereby to provide an exceptionally rich resource about "difficult" words within small-machine memory for on-line access to spelling, definition, and pronunciation. Configuring the lexicon pyramidally by frequency, including all words of high frequency, seems an inevitable model to us now, but it is of course a kind of historical accident.

As many of these comments already make clear, even if one resolves to work within the linguistic bounds of the ordinary print dictionary, there are differences in the demands placed upon the dictionary by print applications and those arising out of electronic applications. It is a matter of judgment or taste for the print lexicographer not to include geographic and biographic terms in the lexicon, but the electronic lexicographer does not have that latitude.

Access to on-line dictionaries can be by the standard alphabetic means or by well-developed phonetic algorithms (which solve the conundrum of needing to know spelling before being able to find spelling) or by definition (the reverse dictionary). As electronic citation for words and senses is done on the basis of machine scans of print-composition tapes and even of voice scans, then sensitive subject coding should permit the development of lexicons tailored to the user profile, with attendant benefits in comprehensiveness and economy of memory. One can conceive of dictionaries that monitor their own use and respond by offering only unkown information to the individual user.

The dictionary that contains synonymy is a resource in the construction of electronic synonym generators, of which there is at least one model that returns synonyms in the inflections of the source words, including phrasal synonyms, taking precise account of all irregularities in doing so. Presentation of synonyms is useful for "knowledge workers" but not for clerical workers.

If usage information is included in the dictionary, then it is deliverable as a discrete electronic product. The most direct key to specific usage guidance is by "trigger" words or phrases that call up guidance information for the operator, but much more sophisticated implementations are possible when programming addresses grammar and syntax.

In large-system management, where accuracy of alpha data is a consideration, the machine dictionary can be the base or one of the bases for verification and correction of data streams in communication or of stored data. What I have called the complete interrelational dictionary--fully coded to reflect the range of significant linguistic information--will serve as the base for retrieving information by meaning rather than mechanics.