THE DICTIONARY SERVER

Martin Kay
Intelligent Systems Laboratory
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, California 94304, USA

The term "machine-readable dictionary" can clearly be taken in two ways. In its stronger and better established interpretation, it presumably refers to dictionaries intended for machine consumption and use as in a language processing system of some sort. In a somewhat weaker sense, it has to do with dictionaries intended for human consumption, but through the intermediary of a machine. Ideally, of course, the two enterprises would be conflated, material from a single basic store of lexical information being furnished to different clients in different forms. Such a conflation would, if it could be accomplished, be beneficial to all parties. Certainly human users could surely benefit from some of the processes that the machine-oriented information in a machine-readable dictionary usually makes available. They can profit even more from other processes specifically oriented to the human user but which have not yet received much attention. For these reasons, I believe that machine-readable dictionaries should, and probably soon will, come to replace traditional book-form dictionaries fairly soon. I do not have in mind machine-readable dictionaries that single users load into their personal machines so much as centralized services to which individual clients subscribe.

I have spend a considerable proportion of the last two years designing and implementing a "dictionary server." This is a computer with a large dictionary in its file system, specifically the **American Heritage Dictionary** (for the use of which we are indebted to its publisher, Houghton-Mifflin Company), together with a variety for indices for giving rapid access to it. The machine is connected thorugh a packet-switching network to a large number of other computers and work stations. Through a mechanism known as "remote procedure call" (RPC), developed concurrently with the dictionary server, a program running on any of these other machines can execute one of a number of procedures "exported" by the server, causing the corresponding procedure to be executed in the server and the result returned to the client as though it had all happened in the same machine. The client reaps several benefits from this mode of operation. First, he does not have to provide storage for this very considerable body of data, nor the time necessary to operate on it. Second, by consigning its care to others, he can profit from regular maintenance and improvements resulting from experience with a large community of users. Less obvious, though perhaps more important than these advantages, is the fact that he can hope to profit from the sophisticated and specialized processing methods available at the central location.

The server I have built represents only a few steps towards the one that would provide the richness of service I can easily imagine. Among its present capabilities are the following: a client can discover if a sequence of letters constitutes a word recognized by the dictionary even though it is presented in an inflectional variant not actually stored in the dictionary. The methods used to accomplished this have sound theoretical bases and generalize across a wide range of language types so that languages with much richer morphological structures than English are provided for. A client can consult the dictionary for the spelling of a word by presenting it with candidate spellings. The server is able to locate entries that could be pronounced in the same way, or ways, as the candidate. It presents these to the client in order of decreasing similarity to the candidate. If the client has difficulty recognizing the appropriate one, he can have the associated definitions presented to him. Definitions, etymologies, synonyms, and so forth can be obtained in a variety of ways. The server undertakes the most onerous procedures that must be carried out by a spelling correction program, namely those that relate putatively misspelled words to actual words into which they can be transformed as a result of the kinds of error commonly made by typists.

These facilities are relatively easy to provide, using as the data base, a machine readable version of a standard dictionary. A dictionary designed specifically for use through a dictionary server could do a great deal more. For example it could present a client with several different perspectives on a semantic field so as to provide a means of finding "le mot juste," that is on the tip of the writer's tongue. This is the function that Roget designed his thesaurus to fill and I believe it is such a device as the dictionary server that will provide the first possibility of doing better than he. What stands in the way of dictionary services of far greater utility than even the largest currently available books is not technological inadequacies, or even shortcomings of linguistic or lexicological theory, so much as the courage and foresight to invest in lexicographic data bases of radical design.