THE ELECTRONIC SPELLER

FAITH W. ECKLER
Morristown, New Jersey

I recently bought a new typewriter--one of those super-duper electronic marvels which does everything but slice bread. Among its many features is a built-in 50,000-word dictionary against which I can check the spelling of words, either as I type them or after I have entered a whole text into memory. When I misspell a word and press the space bar, the machine will beep angrily at me and I have two options: either I can press the space bar again, in which case the machine will reluctantly type the word as I spelled it (probably laughing quietly to itself at my idiocy), or I can ask the machine to search its dictionary and suggest words that I might have been trying to write.

This seemed like a wonderful device until I spent an afternoon finding out what it can and can't do. To begin with, it can't read my mind. If, perchance, I choose the wrong (but nonetheless a legitimate) word, the machine will not tell me. Nor will it signal if I misspell the word I really wanted, but wind up with a legitimate word anyway. I guess I shouldn't expect it to be that smart.

In fact, it's not really so awfully smart. When making suggestions as to the word I might have wanted, it can't handle the possibility that I may have typed two letters wrong. For example, if I type philippines, it will tell me that I should have capitalized the word. But if I write Phillipines, it throws up its hands in despair and has no suggestions as to what I should have written. The only situation in which it can handle two, three, or even four errors is when I fail to capitalize, leave out a letter, and omit one or more periods. For example, if I write ua, one of its suggestions will be U.S.A.

The machine can't do anything about compound words, either to tell me that back bone should be written as a single word, or that semi-automatic should be written with a hyphen. In fact, it seems to have no hyphenated words in its vocabulary.

I wondered about the steps the machine takes to search its data base for suggestions of alternate words. After some trial and error I established the following order of questions the machine asks itself:

1. Should the word be capitalized?
2. Are two adjacent letters transposed?
3. Is one letter in any position wrong? (If so, the machine will run through all its available possibilities, except for abbreviations, in alphabetical order.)

4. Is one letter wrong and the word is an abbreviation?
5. Has one letter been left out? (The need to double a letter seems to be a special case of this question.)
6. Is there one letter too many?

This can be demonstrated by the following sequence which the machine offered me when I typed apr:

1. Apr (should be capitalized?)
2. par (two letters reversed?)
3. air, ape, apt (one letter wrong?)
4. C.P.R. (one letter wrong, abbreviation?)
5. --- (letter left out?) (Aper is not in the machine's vocabulary)
6. AR, PR, A.R., P.R. (too many letters?)

I was curious as to the contents of the machine's dictionary and soon convinced myself that it was almost certainly specially compiled. In addition to its corpus of dictionary words, it apparently contains several other lists. Included is a gazetteer which seems to contain most of the major nations of the world. Of the first 55 countries listed in the gazetteer section of the Merriam-Webster Pocket Dictionary, 1974 edition, all but 15 are included in the machine's dictionary with, surprisingly, many inflected forms such as Argentinian or Ecuadorian.

Also included is a list of proper names, about which I have made certain inferences. Checking Leslie Dunkling's First Names First for a list of the most popular names among boys in the US in 1975, I find that all but three are in the machine. The omissions are Ryan (23), Douglas (36), and Keith (45), although Shane (38), Chad (43), and Bradley (47) are included. Checking Dunkling's list of names popular in 1900, I find none included in the typewriter's data base that was not on the 1975 list except those which, in other contexts, would be words: Frank, Harry, Earl, Ray, Jack, etc. From this I conclude that a fairly recent list of popular first names has been added to the machine's vocabulary. The machine was not so good with Dunkling's list of girls' names popular in 1975; nearly one-third were missing. Of course the machine contains more names than simply the top 50, and here there are some inconsistencies: Frances is in, Francis is not. The list of words recognized by the machine includes a few common surnames that are not dictionary words (Johnson, Williams, Jones, Davis, Anderson, Wilson, Harris, Taylor, Moore, Thompson, Jackson, Clark, Roberts, Lewis, and Allen).

There is also apparently a fairly extensive corpus of abbreviations and initialisms whose full extent and source I have not yet established. All the two-letter postal state-name abbreviations are in, as well as the longer abbreviations in use earlier: Tex., Minn., Cal., Nev., etc. (the machine will insist that you include the period). Other abbreviations include C.P.R. and mpg.

The Greek alphabet is out, save for those letters used in other connections: beta, delta, and iota, although I would think that gamma (as in gamma rays) is at least as common as beta. In fact,

there appear to be no foreign words in the machine's vocabulary unless they have become standardized English, such as **corpus**.

With only 50,000 words, and considerable space taken up with gazetteer, proper names, and abbreviations, one would expect the regular dictionary to consist of fairly pedestrian words. Consider my surprise, therefore, to find **preternatural** and **infrastructure** included. There are some curious omissions, too: **formica, newcomer, Kleenex,** and **Xerox** (capitalized or not). **Acetate** is in; acetone is out. **Compute** is in; **permute** is not, nor is **actuary** or **minuet.** But if this drives you to despair, cheer up; **barbiturate, heroin,** cocaine, and **marijuana** are all in. **Hell** and damn are in, but three other famous four-letter words are not.

I asked myself, "What is the most common word--as defined by Kucera and Francis's Computational Analysis of Present-Day American English--which is not in the built-in dictionary?" Without too much difficulty I determined this to be **negro**, which Kucera and Francis found 104 times in one million words of 1962 text. Its absence from the machine probably reflects a modern cultural taboo; Kucera and Francis's corpus is 26 years old, and negro is more rarely encountered today. It appears impossible to determine what is the rarest word which the machine does recognize. Neither **preternatural** nor **infrastructure** occur in Kucera and Francis at all, and there must be many more such words.

One other curiosity: the machine has to assume that most every word could be capitalized, such as when it comes at the beginning of a sentence. Therefore, Catholic and Episcopal--both acceptable lower case words--are in; Methodist, Presbyterian, and Baptist are out. On the other hand, words which the machine knows only in their capitalized form will be rejected if typed in lower case, even if such a usage exists: e.g., **german**.

So I conclude that the spell-check feature is a potentially useful tool; it can tell me whether **recommend** has two **m**'s and one **c**, or the other way around. But its vocabulary is more limited than mine, and it will not guarantee totally accurate spelling.

And now a challenge for the reader. In the course of typing this article, the machine found four words that were not in its vocabulary. Can you find them? Exclude all my examples, all proper names, and all hyphenated words. There are still four "ringers".