

A Brief Account of Spell Checking as Developed by Houghton Mifflin Company

By: Howard Webber, March, 2007

Author Sketch: Manager, Advanced Development, Office Systems Group,
Digital Equipment Corporation and Publisher, Houghton Mifflin
Reference Division.

Once you begin processing text, computer-based spell checking is the kind of application that is so irresistible that there are likely to have been many "first" implementations and many "first" implementers. It was easy enough to develop a routine that would flag exceptions to a stored dictionary, and many people tried it.

But truly effective spell checking requires a sophistication about natural language that was in the early days not so common in the computer science community. Houghton Mifflin, as publisher of one of the great American dictionaries, *The American Heritage Dictionary*, the print composition of which was driven by a lexical database, began to realize in the early nineteen eighties, as it prepared new editions of this standard reference book, that natural language processing could assist it in ensuring that the lexicon was as accurate as possible and reflective of the new and not entirely welcome standard set by editor Philip Babcock Gove in Merriam-Webster's *Third New International Dictionary, Unabridged*, that dictionaries should reflect what people were actually saying and not an abstract, inherited standard. That latter objective required much more exacting measurement than had been done before.

Since time immemorial, that is, since James Murray and his staff put together the Oxford English Dictionary—the famous O.E.D.—in the latter part of the nineteenth and early part of the twentieth centuries, dictionaries had been compiled on the basis of a process known as "reading and marking." Dictionary publishers would employ retired clergy, for example, or schoolteachers on a summer break to notice new words or new senses of old words and to submit slips recording the words in context to the dictionary editors as documentation. (Of course, for Murray the challenge was even greater because the Oxford wanted to be a "historical" dictionary, representing obsolete words and words still in use with the dates of the first appearance for all entries and with citations.)

When I came to Houghton Mifflin in the early nineteen eighties, the commitment had been made to use the computer actually to document American English as it was written at the time. How could that be done? An alliance was formed with Dr. Henry Kučera of Brown University, who as early as 1962 was already teaching a course in computational linguistics there. He was one of the principals of the U.S. Office of Education's Standard Corpus of American English Project which resulted in the creation of the well-regulated million-word "Brown Corpus," on which was based his *Computational Analysis of Present-Day American English* (with W. Nelson Francis) and *Frequency Analysis of English Usage* (also with Francis). This work offered the opportunity of developing a lexicon on scientific principles rather than the accident of personal recognition of new words, forms, and senses. From the perspective of

dictionary-making, this resource allowed the lexicographers to determine what words were being used and how frequently—but as an equal benefit, what words had passed out of use. It was manifestly not possible to ask the readers-and-markers what words they were no longer using, yet culling the lexicon in this way was very important in order to produce an accurate contemporary dictionary. It was also important in producing a lexicon of controlled size, to accommodate the memory limitations of early computers but also to avoid the rare words that most users probably didn't mean.

I should note that "collegiate" dictionaries, like *American Heritage* or the *Merriam-Webster Collegiate*, contain only a limited subset of available words, perhaps 150,000 of them, whereas if you include scientific, industrial, technical, commercial, and dialect words as well as non-English borrowings, there may be as many as 4,000,000 words in the "American English" language (my guess, which, as any lexicographer will tell you, is a real shot in the dark). So the Kučera work enabled Houghton to determine which of the large universe of available words should be included in a print dictionary, and, moreover, provided the technology by which the lexicon could be constantly renewed as the language changes.

It seemed to be an easy jump to a dictionary that would drive spell checking, and in fact the first versions of the Houghton checker were very useful, particularly in spotting errors on screen, where somehow human proofing was not as successful as it was in print, even if the human could spell. But the electronic spell checker, it became clear, needed to include proper nouns, including commercial names, that were not commonly included in the print dictionary.

But under the tutelage of Dr. Kučera and his company, Language Systems, Inc., Houghton came to understand that if it was possible to do a basic amount of parsing for the purpose of disambiguation, you might sometimes—not always!—be able to help people choose among homophones (*all* and *awl*; *to*, *too*, and *two*; *rite*, *write*, *right*, and *wright*) and homonyms (*pool* of water and *pool* the game), all this under the rubric of spell checking.

One thing led to another. Could you develop spell checkers for foreign languages? Yes, you could. Could you tune spell checkers so as to respond to the needs of non-native English speakers? Yes, and Chinese and Japanese working in English would find your product highly desirable. If you were beginning to parse automatically, how about combined spelling and grammar correction? If you knew a lot about the relationships between phonemes and graphemes, how about speech-to-print products? Or, turning them inside out, how about speech generation? And behind all of this, how about intelligent (in the beginning, we used that word a lot) retrieval, because wasn't retrieval basically a matter of filtering text for defined words or combinations of words? Well, as we've discovered, only in part.

Houghton's spelling correction was licensed to stand-alone word-processing systems like Lanier and to Microsoft for Word, and to many other such users. But essentially, the language-processing opportunities began tumbling over one another, and in some sense they overwhelmed the rather conventional publishing environment in which Houghton was accustomed to working. In 1994 Houghton divested itself of its language-processing products in favor of a new company, Inso.

In the length of time and after a large number of scattershot acquisitions of targeted language-related products and some questionable financial transactions, Inso bought

yet another company, Electronic Book Technologies, Inc., in a kind of merger deal, and the combined company sold its language-processing technology to Lernout & Hauspie, a respected Belgian company (whose investors included Microsoft and Intel), in favor of an emphasis on electronic publishing and document conversion. But the respect was not entirely deserved, and L&H was discovered to have engaged in serious fraud in the effort to brighten up its financials.

I won't go on with the story but will only call attention to this tale as quite typical of the Wild West days of early computation when legitimate promise was so often outstripped by fantasy and woeful underestimation of resources and time required to reach stable product status. It was just the same with James Murray and the O.E.D., who to the agonized protests of Oxford University Press, completely underestimated the time and money required to bring his project to completion and died before he was able to see his life's work in print. "In the interests of the Dictionary," he wrote to a friend about his relationships with the Press, "I have desired not to raise the question of time, or any other, until we get the work fairly launched." He was at the point thinking an appalling sixteen or seventeen years, but in fact the task took far longer.