The Paper Killer

I NEVER BEFORE BELIEVED THE CLAIMS OF COMpanies that sell optical character recognition (OCR) software—those programs that turn scans of printed pages into editable text. That's because I know how to multiply. When the companies claimed "99 percent accuracy," I translated that to roughly one error on every line. And that, in my opinion, was unacceptable. ■ Then last fall, I had a revelatory experience. I was trying to sell an old book of dolls on eBay, so I scanned a page and told Microsoft Office Document Scanning to save it as an image. When the program prompted me for a file name, the name it suggested corresponded to the scanned page's headline.

Funny, I didn't remember typing in that headline.

On a whim, I saved the page as a Microsoft Word document. Then I opened the file and compared it to the original. There was not a single error in five paragraphs of text. Optical character recognition software has done a lot of growing up in the past ten years. A lot of people don't realize just how good the technology has become because they base their impressions on the free software that comes with scanners rather than on professional-grade software, which costs hundreds of dollars. But after I wrote last year about how I was scanning all of my old documents into PDF files *(see "Slaying the Paper Dragon,"* TR *October 2003)*, a publicist at Abbyy USA sent me the company's FineReader 7.0 corporate edition.

My test was simple: I fed the software my bank statement. FineReader spent 15 seconds scanning the page and another 30 turning the image into text.

Modern character recognition systems use a variety of mathematical techniques to perform their magic. Imaging algorithms remove speckles and rotate the page so that it's "straight." Then a series of algorithms separates out each glyph, determines the likelihood that any glyph is a particular letter, and consults a dictionary to come up with a probable word. The software can also decide to accept a word that's not in the dictionary, if the image looks good and there are no obvious close matches.

The cost of optical character recognition software is far lower than that of the alternative: hiring a typist.

Optical character recognition has made such strides in recent years for much the same reason that speech recognition and machine translation systems have: fast computers with massive memories allow software to evaluate many alternative hypotheses when it's trying to recognize words. Programs can consult statistical models created by analyzing millions of pages of text and pick hypotheses that have the greatest chance of being correct. And the explosion of the Web has provided millions of pages of text from which to compile the statistics.

Because these techniques have become widespread, I wasn't surprised that FineReader could correctly recognize the legalese that my bank printed at the bottom of the statement. What surprised me was that it could handle the rest of the page: it recognized my street address, the names of companies that had received electronic payments, and even the dollars and cents of every returned check. The reason this is so amazing is that FineReader can't use dictionary lookups to improve its accuracy on these items; it has to get them right based on image processing alone.

I also experimented with OmniPage Pro, a similar program from ScanSoft. Like FineReader, OmniPage allows you to manually review and correct errors, spell-check the OCRed document, and save the resulting scan in a variety of formats—including Microsoft Word, HTML, and PDF.

PDF is the best for my purposes, because it's the only format that allows me to save both the original page and the OCRed text in a single file. I need the original image in case I get audited. But a computer can't search for individual words in an image; for that you need text. A PDF file made from a Microsoft Word document can be searched because the text has already been entered letter by letter; a scanned-in document, however, is unsearchable unless you run OCR.

Based on my OCR experience, I'm now going through my last seven years of bank statements. Each year gets loaded into my sheet feeder, scanned, and OCRed into a single file. If I ever need to find something, I'll just perform a computer search for it; in my tests, searching is faster than manually paging through the paper statements. I've also started scanning my deceased father-in-law's poetry and other writings from the 1960s and '70s. I always intended to put this material online—and with OCRed text, Google will find it. The modest cost of the optical character recognition software is far, far, lower than that of the alternative: hiring a typist.

And optical character recognition is going to get even better within the next few years. Techniques under development will use grammatical models of English and other languages to disambiguate words that are visually similar but grammatically different—words like "bottom" and "button." Should I want, I'll be able to open those PDFs and reprocess them.

I'm slowly getting rid of the paper dragon in my basement. At this rate, I may really be paperless by 2005.

Simson Garfinkel is an incurable gadgeteer, an entrepreneur, and the author of 12 books on information technology and its impact.