[*Translating and the computer 12: Applying Technology to the Translation Process*. Proceedings of a conference … 8-9 November 1990, ed. Catriona Picken (London: Aslib, 1991)]

Document conversion

Peter Laurie Southdata Ltd

and David Hickman Real Time Implementations

David Hickman:
Peter Laurie and I will now tell you something about the problems of document conversion. Peter will cover general problem areas, and I will follow up with possible solutions to those problems that are commercially available in current software packages.

Peter Laurie:
You thought you'd heard the last of the bad news, but now we are talking about file conversion. You've gone through the horrid process of OCR, and you've got some kind of representation on your computer, and, of course, remember you're not just talking about a computer, you're talking about a computer with quite elaborate word processing software in it, which allows you to deal with text represented in some sort of way on the screen. Now we want to get it to a different machine and possibly into a different software package. What are the problems?

The first one, which I know most about because we do it every day, is actually the simplest: you take your file, you stick it on a floppy disc and then you want to turn it into another floppy disc for another computer; it's a very easy thing to do, you just send us the disc and we'll convert it. Of course nowadays so many machines use IBM discs that this stopped being the crucial problem it was four or five years ago when there were something like two or three hundred difference disc formats in use. That has all abated. But you might have a tape from an ICL mainframe and you want it on a little Amstrad word processor. Yes, it can be done. The really tricky problem, is the representation of different characters. Now we considered that just now in the case of a chemical formula which can be printed by a rather elaborate and obscure code in my word processor, but which would probably be totally different to the word processor codes in yours. So we need to have some kind of conversion of these codes.

Of course we could have a package which would convert from one word processor to another, but there would have to be hundreds of different combinations, and software manufacturers keep producing new versions. I suppose the long term solution is to have some kind of standard code which rises above this petty morass and says in nice unambiguous standard ways that we want to have a one up and a five down, whatever it is. That sort of thing does exist, I understand, in the Pentagon, in Washington.

David Hickman:

I'm going to outline the problems of document conversion and basically put forward and reiterate an idea that was mooted at the conference two years ago. I'll start with a quote that says "the ideal system for use by professional translators is one which permits the full exchange of documents between different systems and different software packages without the need for special file translation programs or utilities". That was a quote from T + C 10, from David Jackson who is the managing director of Vuman Computer Systems. I didn't feel that Vuman had really tackled the problem, which is actually the technical root of the character transfer problem itself. The inclusion of special characters generated by particular word processors included in their output files, generates special symbols for foreign characters, and means that the two software packages involved must be absolutely one to one identical i.e. everybody uses WordPerfect 5 or whatever. This clearly isn't the case, because of the diversity of scale of the translation being performed by people in this room, from one-man freelance translators to huge corporate translation departments with vast resources. As already been indicated, there are thousands of types of computers around these days – the clones, the Amstrads. To produce different variants of software packages for each one of those is not cost effective. We can see why the major corporations don't invest time and money in that.

Now, what still hasn't been achieved over the past two or three years, is an agreement on the basic code pages which define the character sets available within the word processor itself. The first code page produced and defined by IBM was called code page 437. Again, as has been indicated and commented upon many times, it makes you wonder what IBM were actually thinking about when they developed the PC. Clearly they played a lot of cards and there are a lot of boxes but apart from that there's not a lot of merit in the code page itself.

In recent versions of MS-DOS, and I stick with MS-DOS, as opposed to the other commercial operating systems that are available, there has been an attempt to reduce the scientific character content and obviously to bring in more foreign language characters which would be useful to translators. The only problem with that is, while code page 850 works perfectly well in certain printers that IBM produce, and you've got all those wonderful commands in MS-DOS to specify code page 850 for your printer and your EGA screen, if you haven't got that particular type of hardware and software, you really can't make much use of code page 850, which probably explains why it hasn't exactly become the universal standard.

What I'm saying is that the problem is basically technical. It doesn't strike me that you have to be a great genius to work out the fact that you can look up one character in one table, and translate it to a character in another table. In computing terms that's known as a simple, straight look-up. But we can't see that logic being applied to the packages that are currently available. In terms of the problems that face you now, and the problems that you have to resolve when transferring documents between companies, between your sub-contractors and wherever else, there are a number of packages available which will perform document conversion. I can't say that I am absolutely familiar with all of them; probably you have more experience of them than me, and it might be useful for instance in the question and answer session to hear your views, if you've used any of the packages that I've indicated.

Now, I'm a great WordPerfect fan. Some of you may not be. But from the point of view of tackling multinational character sets and different symbolic and math and scientific sets WordPerfect Corporation have actually adopted the right logic in solving the problem. Unfortunately they seem totally unwilling to give or divulge that knowledge to anybody else. There are a number of character sets available for selection within Word Perfect. The base set, set zero, contains the normal 437 basis of 12 codes. There are two multinational sets, a maths and scientific set, extensible and non-extensible, Greek, Hebrew, Cyrillic and Japanese (two types). Now the inclusion and the generation of these particular character sets is quite an easy operation in Word Perfect. You just choose the character set and enter the code corresponding to the character you wish to generate.

Each word processor has its own way of saying "I wish to generate these codes" and instructing the target printer or the screen to actually display those for you. Most word processors don't actually allow you to see the character you've generated, it merely shows as a blob on your screen, so you don't actually know if you've actually generated the correct code.

I believe that a generic solution could be based on the substitution of token characters for individual characters within those character sets themselves. If the token table could be adopted across the board by word processor manufacturers, then we wouldn't need switch between multinational character sets, because the one token would uniquely identify the character. And it needn't necessarily be an ASCII character.

I'd just like to give you a brief description of a number of packages which are available that actually claim to perform document conversion between some 20 or 30 of the available word processors in the market place. The first one is produced by a firm call Migent known as Word For Word. This is essentially a word processing conversion program and only uses IBM code page 850 for its multinational character sets. Any other characters therefore must be inserted in the native format of the word processor or a token inserted and then the global search and replace option used; I'm sure that's something you've all done many times to put in the Polish diacritics-you can actually put a substitute character in there and then every time you see that, you insert this particular Polish character. The advantage of this package is that it will cover a large number of formats, both with importing and exporting.

And the second package was produced by a firm called Formscan. This actually is a set of OCR scanning software and conversion utilities all put together in one package, called Omni Page, which is both conversion and OCR system, but is only a partial solution to the general problem, because it requires you to have an original document for scanning in the first place, and we know the problems associated with original documents that were highlighted by Peter Laurie earlier. This product seems to be extremely useful for the technical translator who may work with a mixture of both text and graphics. Conversion of scan text into nearly all the major word processors' format is provided for and just as a little sales plug for them, the software is currently being marketed by Formscan in conjunction with Hewlett Packard, who actually produce an excellent scanner called the ScanJet Plus, which is on special offer in case anybody's interested.

The third package is one probably which I think will be familiar to a lot of you who use the Amstrad PCW – Loco Script, the original PCW Word Processor which has been developed now for running under MS-DOS and is called Loco Script PC. This popular word processor is now available on PC and offers good multinational set support and a wide variety of printer drivers. Languages currently supported are Western European, Eastern European, Ancient Greek and Cyrillic, with customization kits for Arabic and Hebrew. I think that's a very important feature of the packaging; you can customize it by bolt-on software updates to the package. Quite a recommendation, in my view.

To summarize, the problems that existed four or five years ago still beset you all and still aren't really resolved by the software packages currently available. I can only hope that in the coming twelve months before the next conference, you will see some movement on that front.