COUNTING WORDS

I once commissioned a 2,000 word article for QL Today and when the author submitted it he proudly announced that it was exactly 2,000 words, no more and no less. My sense of mischief took over and to spike his gun I ran the article through several PC and QL programs with the following result:

Microsoft Word:	*2,000*
Microsoft Works:	*1,995*
Lotus WordPro:	*1,995*
Quill:	*1,995*
Perfection:	*2,002*
Text87:	*2,001*
Style-Check:	*2,003*

It did not surprise me that different word processors gave different word counts for this document. What did surprise me is that the difference between the highest and lowest word count was only 8 words, and that three word processors came out with the same value of 1,995 words.

Word counts in word processors are estimates - albeit in most cases good estimates. It is well nigh impossible to make an accurate word count by computer and I suspect, although I have not researched this, that word counts of technical documents may show the greatest differences between word processors.

I learnt a lot about the problems of counting words when I wrote Style-Check. Much of the work of style checkers is done by statistics and thus you need accurate ways of calculating the number of words, the number of syllables and the length of words in a document. It is far more difficult than first impressions suggest.

The simplest and crudest way of making a word count is to count the number of spaces in a document. However this will only give an approximate result as it does not take into account the idiosyncrasies of different writers. Sometimes there will be more than one space between words. In the early days, when fixed font widths were the norm, some word processors used spaces for tabs. There are writers who do not leave a space between sentences or after a punctuation mark. And in most word processors there is a Line Feed instead of a space at the end of each line. You thus have to count spaces, punctuation marks and Line Feeds and then write routines to check that you have not double counted.

There can also be a problem with both soft and hard hyphens. Most of us would say that 'co-opt' is one word, but is 'ink-well' one word or two? It would be an onerous job to write a routine to distinguish between the two, and instead you choose either one or the other for all hyphens.

Most people quickly suggest this as one of the reasons for differences in the word count between word processors, but there are many others. “Can't” is short for “cannot” which is one word, but “don't” is short for "do not” which is two. Again it would be difficult to write a routine to distinguish between the two. If I am a VIP and I have just got angry in public and told someone to "bugger off' it might get reported in the papers. Some journalists would use the term itself, which has two words, but others would prefer the form 'b*gg*r off', which would be four words in a word count unless you wrote a routine for this.

Nearer home there are the computer file names. Is "program_bas" or "program.bas" one word or two? Almost certainly two words unless we wanted to write another complicated exception routine.

Numbers are a big problem. Should they be treated as words? “4" is simple as it is pronounced "four", but “444" is pronounced 'four hundred and forty four". Is it one word or five words? Publishers use the word count to estimate the space an article needs and broadcasters to estimate the length of a talk. In a piece on, say, economics containing numerous statistics there could be a big difference.

Decimals are an even bigger problem. "4.4” would come out as two words unless we write another routine to distinguish between decimal points and full stops. And what about dates? The QL was launched on 12.1.1984, which is three words unless you have an exception routine.

I learnt something of the limitations of counting statistics by computer when Style-Check assessed an academic paper I had written as being at the reading level of a primary school child. The article was a research study on the growth of criminal sophistication in a group of adolescent delinquents and contained a detailed diary of their numerous court appearances with dates in the form of 12.1.1984. Style-Check assumed that the piece contained many sentences that were both short and had no long words and thus was suitable for young children.

Academic papers can give many problems especially if they contain mathematical, chemical and other scientific formulae or Greek letters. Is boric acid, $B(OH)_3$, one word or three?

One thing that I did not have to worry about when I wrote Style-Check was smileys. Should smileys be counted as a word or not? Most western smileys are made up from punctuation marks and will not appear in a word count. The occasional one - (: - 0) - will.

In summary you have to be sceptical about the word count that your word processor produces, and some texts should be treated with greater scepticism than others.