

## SOME DICTIONARY COMPILATION PROBLEMS

Some years ago I sent a Just Words! German dictionary to a colleague for evaluation. He replied:

*'I think this is an unpleasant message for you.'*

In fact the email did not upset me as much as he feared, because I had been expecting that sort of reply. Unbeknown to him, the material I had asked him to comment on was not entirely my own work.

The German gave me a detailed and technical reply. He found the word list was a mixture of old and new spellings: had some missing declensions and conjugations: and some spelling errors and doubtful words. His general conclusion was that it was probably OK for non-Germans to use but had too many errors for serious native German users.

What I had not told him was that I had used the German dictionary of a well known PC word processor to check the accuracy of my word list. The German's observations confirmed my impressions that PC spell checking dictionaries are not 100% accurate. In that word processor's English dictionary I had found many missing plurals or strange plural forms. There were few personal names and inconsistencies in the inclusion of place names. Most verbs ending in 'ise' or 'ize', used the 'ise' form even though the Oxford English Dictionary gives the 'ize' as the preferred form for some.

I am not the first QL user to comment on the quality of PC dictionaries. Over 10 years ago Digital Precision wrote in their Spellchecker manual,

*'We have not used computer methods for generating words (most words ending in ed' can also take an 's') - this short-cut leaves out many legal words and includes hosts of illegal words (in one case we estimated that about 50% of a supplied dictionary was illegal). It was our object to provide you with good dictionaries (ones which have all the common - and most uncommon - legal words and no illegal ones), not just big ones. In this area good implies big, but the reverse is not necessarily true.'*

Those of us who have written dictionaries for QL spell checkers can be forgiven a degree of Schadenfreude when we discover PC spell checkers are not 100% accurate. We know, hard as we try to be accurate, that errors creep into our word lists. Even when we know the correct spellings an occasional typing error occurs that we do not detect, because our brains are programmed not to notice an error if we understand what is meant. Can you spot, without stopping to examine them, the difference between the correct and incorrect spelling of the Dutch words:

*'eigenlijk' and 'eigelijk'  
'burgelijk' and 'burgerlijk'  
'onmiddelijk' and 'onmiddellijk'?*

Often when we add a new word to a dictionary we include all its declensions or conjugations, and sometimes we create inaccuracies like 'unique, uniquer, unquest'. This is almost inevitable when we attempt to automate the process. In English we add an 's' to make a plural so in an automated list the plural of 'child' becomes 'childs' and of 'children' 'childrens'. One QL Dutch dictionary produced in this way gave the word 'rectangular' a comparative 'more rectangular' and a superlative 'most rectangular'.

Then there are the foreign words that can creep into any language. A couple of paragraphs ago I used the German word 'Schadenfreude'. This is now permissible to use in English, although the writer who uses it risks being accused of intellectual snobbery,

because it conveys a concept that cannot readily be expressed in English. Should it be included in an English spell checking dictionary? And should we include our personal words? As someone put it to me about the content of one QTYP dictionary, 'I never knew that 'Tony' and 'Tebby' were Italian words'. (Tony Tebby is a well known figure in the QL community.)

Another difficult area is the words that may cause offence. In general I do not believe it is my job to censor a word list for spell checking, but I am not entirely consistent. Usually I do not object to slang, swear words or sexual terms, but I have a marked reluctance to include racist or other discriminatory words. Another exception is where a word list is likely to be used by children in, say, a word game, when I impose a higher level of censorship.

Many of the dictionaries that were available for QL spellcheckers were substandard and some years ago Just Words! systematically revised and corrected these. They are available on our Dictionaries page. However the greater the size of the dictionary the greater the chance of some inaccuracies.