Using corpora to develop materials on collocation*

As someone who writes materials for Cambridge University Press, I am able to use their enormous English language corpus. This continuously growing corpus is, I believe, the largest in existence with currently around 800 million words of written English and 40 million words of spoken English

It is, of course, an amazing tool and one that pre-1990s' textbook and dictionary writers could only have dreamt of. But how can the ordinary writer, rather than the specialist lexicographer, actually handle this powerful tool?

Most recently I've been working with Mike McCarthy on some materials which deal with collocations. The first and most difficult task was to select the collocations that we would present and practise in the book. There are just so many collocations in the English language. Which ones would it be most useful for us to work on with our target upper intermediate learners? How on earth could we narrow things down in a reasonably principled manner?

You might imagine that it would be easy to ask the corpus to do a frequency count. But this is not, of course, as straightforward as doing a frequency count for individual words. Frequency information can be collected about clusters of words but it is clearly necessary to cut out those little grammar words which do not tell us anything about key collocating words. Coping with this is certainly not something that we could easily do ourselves but the specialist corpus team at the Press were able to help us. They did some initial work on three-word clusters cutting out just *a, on* and *of* and working initially with just one section of the corpus. Here are the top 20 results in rank order, reading down the columns. It should not be too difficult to work out which section of the corpus they were working with!

this is what
lord your god
and i will
lord said to
he said to
said to him
you will be
said to them
i tell you
and you will
will not be
it will be
said to moses
i will not
you do not
there is no
they will be
will know that
i will give
at that time

While this information is clearly interesting it does not tell us anything that has relevance for a collocations practice book, even if we were wanting to include a unit on Biblical collocations. Far more of those little grammar words have to be excluded before we get down to significant collocation information.

Using part of a British newspaper corpus, the techie wizards at CUP were able to exclude more such grammar words and tried out their new programme on one year's worth of a newspaper corpus. From this process the resulting 20 three-word clusters emerged as the most frequent.

10 per cent
50 per cent
second world war
mr major 's
15 per cent
20 per cent
40 per cent
mrs thatcher 's
prime minister's
30 per cent
25 per cent
exchange rate mechanism
100 per cent
south africa's
premier john major
chancellor norman lamont
hong kong's
60 per cent
's world cup
john major's

Again there are interesting aspects about these data. For example, like an increasing number of people, the computer seems unable to interpret the significance of the apostrophe. But the above list clearly indicates more about the year that the particular corpus selection was taken from than it tells us about collocation. (In case you're wondering it was 1992).

Of course, one of the major difficulties in finding frequent collocations by looking at three word clusters is the fact that the two parts of a collocation are often separate. *Doing homework* or *making mistakes* are collocations that students are likely to need to know but they are more likely to be found in longer and widely varying clusters - *do a lot of homework, do some maths homework, make a terrible mistake, make a number of mistakes*, for example. Using the corpus to get frequency data relating to collocation is clearly no mean feat.

Eventually, by using some kind of magic spell, the corpus wizards were able to come up with an excellent frequency list of collocations for us but it took them some time and in the meantime we carried on selecting collocations in our own way. This was by deciding on the unit titles we wanted for the book and then looking up in the corpus key words relating to that area. For example, we wanted to include a unit on weather, so we then ourselves looked up words like *rain, wind* and *snow* in order to select the collocations that we felt would be most useful for learners. We wanted to have a unit on intensifying adverbs, so we looked up words like *utterly, highly* and *bitterly* to see which adjectives they most frequently collocated with. This process provided us with a lot of very interesting and usable material.

This was despite an initial hiccup when I did not have the parameters of my corpus data set correctly. Instead of using the whole corpus, I had restricted my search to literary sources. This meant that expressions like *shrouded in mist* and *through a mist of tears* emerged as more frequent than *a light mist*. I thought I had made an amazing discovery about English usage until I realised my error.

Having - with the help of the CUP corpus team and through our own less systematic method - managed to select the collocations that we wanted to present in our materials, it then became more straightforward for us to make unaided use of the corpus to find examples of their usage. Having decided, say, that good collocations to deal with in a lesson on describing people's character might include *bear a grudge, painfully shy, a selfish streak* and *an aspect of someone's personality* how could we provide an appropriate example sentence to encapsulate the essence of how each collocation is actually used? Would it best be presented through an example of written or spoken text? Is there anything else that is characteristic of its typical context that it would be helpful to suggest to learners? It's often quite difficult to be sure about the answers to such questions and I found my instinct often let me down. Either I couldn't think of any way of using a specific collocation that felt just right or I got stuck on one idea - enough perhaps for a presentation example but it wouldn't then be helpful or appropriate to reproduce that idea several times in the practice exercises. At the click of a button, the corpus can suggest a whole range of examples to save materials from being too idiosyncratic, too restricted by the limited instincts or experience of an individual writer. It makes a whole world of language accessible in seconds from your own desk.

However, it must be pointed out that all this rich information to be learnt from corpora has already been analysed for us and is available in excellent learners' dictionaries. These, for instance, indicate frequency and highlight collocations through examples and sometimes also through special features. For most of my collocation needs it was - more often than not - not actually essential to go back to the primary source of the corpus.

This can save a massive amount of time. Using the corpus easily becomes like looking something up on the web - it is all too easy to move from quickly checking out one specific item into exploring all sorts of other things that catch your eye or jump into your mind. Before you know it you've gone from *a selfish streak to streaking across a cricket pitch to silly mid on to the silly season* and a morning has passed in a fascinating but not very productive manner. Using the *Cambridge Advanced Learner's Dictionary* allowed me both to access a mediated form of the corpus and to meet my deadlines.

Felicity O'Dell
Cambridge, September 2005

**Article originally published on the Humanising Language Teaching website.*