Helen Myers

How confident are you that a corpus which draws a great deal on parliamentary debate (4.7 million of the 11.5 million contributing to speaking corpus) reliably identifies high frequency words suitable for a GCSE examination?

Emma Marsden

Hi Helen. In these slides here https://resources.ncelp.org/concern/resources/v979v4156?locale=en you can see a comparison of that corpus with another large, general corpus (these are the same slides as those I sent you the other day). See slides 44-50. The corpus is called TENTEN, and used quite a lot in research and for pedagogy. Texts only from internet - so, likely to be more informal. The French version is pulled from many varieties of French language – European, Canadian & African: 5.8 billion words The overlap is 82% with the Routlledge '2,000'. That is, 1,635 of the 2,000 word families from the Lonsdale & LeBras (Routledge) list are also in the top 2,000 words found in this other massive general corpus. This kind of finding is replicated over and over, in different languages. Also, worth bearing in mind that the AOs only need about 1,550 words (as some of the 1700 are the very highly irregulars). So the words chosen could, in the end, be in BOTH large corpora! Even those engaging parliamentary debate need those high frequency words :) (By the way, only about 200-300 of those are the 'function' words. The rest are interesting 'content' words which lend themselves to various different themes and topics.

I think it would be wonderful if people did some research on tweaking that corpus to add more texts and conversations that teenagers engage in. But of course we don't want to just prepare our students for 'teenage speak'! (A corpus of 'text speak' is available and that makes interesting reading!). We want to prepare them for language that is useful across a range of ages, contexts, genres. The evidence suggests that the highest frequency words wouldn't change much. I think that the AOs will of course engage teachers in deciding the words for the 15% words that can be of any frequency. (In fact, some of our resource developers at NCELP found that they didn't want to use the full 15% of words outside the 2,000, when they created some example lists using this 85: 15 balance. See https://resources.ncelp.org/collections/jd472z20c?locale=en. Of course those are just examples lists! The AOs will be providing the lists.

Hope that helps.

Emma