%DIFF is a simple and straightforward effect-size metric for keyness analysis. It indicates the proportion (%) of the difference between the normalised frequencies of a word in two corpora (or sub-corpora). %DIFF is calculated as follows (Gabrielatos & Marchi, 2011):

$$\%\text{DIFF} = \frac{(\text{NF in SC} - \text{NF in RC}) \text{ x } 100}{\text{NF in RC}}$$

NF = normalised frequency
SF = study corpus (corpus 1)
RF = reference corpus (corpus 2)

Please note
However large the %DIFF may be, it also needs to be statistically significant. That is, frequency differences that are below the statistical significance threshold adopted in the study should not be included in the analysis (see here for more information and threshold values).

FAQ

What do you mean by ‘effect size metric’?
It measures the size of the difference the normalised frequency of a word in two corpora, not the statistical significance of that difference.

So what does the LL values show?
They show the level of statistical significance of the difference we have observed. Simply put, statistical significance values (p-values) show the extent to which we can be confident that the frequency difference we have found is dependable.

Why are some %DIFF values negative?
Negative values indicate that the particular word has a higher normalised frequency in the reference corpus (corpus 2).

What is the cut-off value for %DIFF?
There isn’t one. The threshold has to be relative to the resulting range of %DIFF values (which will vary according to the corpora compared). For example, a 50% DIFF is relatively ...
... small, if most values are larger than 200%.
... large, if most values are smaller than 20%.
Please also keep in mind that the cut-off values for statistical significance are also arbitrary. For example, there’s no other reason but consensus that CL researchers are happy with a p≤0.01, but not p≤0.011.

When the reference corpus (corpus 2) has no instances of a word, how do you divide by zero?
All zeros are converted to a infinitesimally small number (0.000000000000000001 – one quadrillionth), which, for practical purposes, is an adequate proxy for zero, while allowing for division by it. Please note that this results in extremely high %DIFF values.

Why are some %DIFF values extremely high?
These high values are due to one of the corpora having zero occurrences of a word. Very large %DIFF values flag up potentially interesting differences. However, the LL score will indicate the extent to

which we can trust this large %DIFF score (e.g. in corpora of roughly the same size, a couple of occurrences in one, and none in the other, won't be statistically significant).

Why not exclude instances of zero occurrence from the comparison?
Excluding such instances may well remove very interesting/useful differences. If 'more occurrences than the other corpus' is interesting, then 'some/many occurrences in one corpus, though none in the other' is even more interesting, because these words can be seen as characterising not only the corpus with non-zero occurrences, but also the corpus with zero occurrences – particularly if these words are related by meaning or use.

Why not use the 'add 1' technique?
(NB. When using this technique, '1' is added to the normalised frequency of a word in *both* corpora, so that zero frequencies are treated as if they are 1.) Because by adding '1' to the frequencies we do not increase both frequencies equally. For example, by adding '1' to a normalised frequency of 50, we increase it by 2%, whereas by adding '1' to a normalised frequency of 5, we increase it by 25%! In other words, this technique skews the results.

What is the difference between 'study corpus' and 'reference corpus'?
The distinction is just one of focus. We can compare two corpora (e.g. A and B) twice:

- A is the study corpus, B is the reference corpus.
- B is the study corpus, A is the reference corpus.

Does the reference corpus need to be larger than the study corpus?
No -- study and reference corpora can be of any size. Corpus size differences are taken into account by statistical significance tests.

Does the reference corpus need to be a general corpus?
It can be, but it doesn't have to. The selection of corpora to compare is only dictated by the research focus and questions/hypotheses.

Please direct further questions to:

Dr Costas Gabrielatos
Senior Lecturer in English Language
Department of English & History
Edge Hill University
Email: Gabrielc@edgehill.ac.uk