Word Usage Comparisons

What do they do? Given two corpora, a outputs a list of words that distinguish the two fields most strongly.

What's so great about them? More than anything else, this tends to be something humanists actually *want*.

What choices do I have to make? Most importantly, the algorithm.

As in topic modelling, the *text* itself. Very few word comparison tools work on more than two *texts*; so frequently you have to decide how to lump many shorter texts together with each other, or do a number of one-to-many comparisons.

Also, what a word is, to begin with (see *tokenization*).

What algorithms are there? Too many, because none of them are especially good.

The simplest choice is the "odds ratio," which gives the probability in one set over another.

Dunning Log likelihood gives a probabilistic version that tends to heavily weight common words.

Mann-Whitney scores look at the relative *ranks* of words in sorted order.

What software should I use? Voyant implements a number of distinctive word tests out-of-the-box. (Not, I believe, including Dunning or Mann-Whitney, though.) Try Z-score for a basic comparison tool.