University of Glasgow
College of Arts
SGSAH/SGSSS: Spring into Methods

Monday 16 April
1400-1530
Professor Marc Alexander

Analysing Corpora: Keyness

Keyness tests "compare the difference between the actual frequencies observed in the corpus (the *observed* frequencies) and the frequencies we would expect if no factor other than chance had been operating to affect the frequencies (the *expected* frequencies). The closer the expected frequencies are to the observed frequencies, the more likely it is that the observed frequencies are a result of chance. On the other hand, the greater the difference between the observed frequencies and the expected frequencies, the more likely it is that the observed frequencies are being influenced by something other than chance." (McEnery and Wilson 2001: 84-85)

Stats and Keyness

95th percentile; 5% level; $p < 0.05$; critical value = 3.84
99th percentile; 1% level; $p < 0.01$; critical value = 6.63
99.9th percentile; 0.1% level; $p < 0.001$; critical value = 10.83
99.99th percentile; 0.01% level; $p < 0.0001$; critical value = 15.13

"There is no agreed-upon 'cut-off' point on what chi-squared or log-likelihood score results in something being defined as a keyword or not. Instead, we can vary our notion of what is 'key' depending on how many keywords we want to examine (which is often constrained by issues of time, money or publishing word counts). In general, the larger the corpora we are examining, the more keywords we are likely to elicit. Some corpus linguists have therefore backgrounded the notion of statistical significance, favouring instead a method of focusing on the 20 (or 50 or 100) keywords that have the strongest keyness score in a corpus." (Baker 2010:26)

Murder and the Reference Corpus

Corpus 1: Masters' thesis about Agatha Christie
Corpus 2: Reference corpus, Wikipedia (one million random sentences from Wikipedia)

The word 'murder':
Corpus 1: 172 occurrences
Reference corpus: 8714 occurrences

What if there should be no difference, proportionally?
Corpus 1: 9.8 occurrences
Reference corpus: 8876.2 occurrences

	Reality	**Prediction**
Corpus 1	172	9.8
Reference Corpus	8714	8876.2

Log-likelihood ratio *for this word in this corpus with this reference corpus*: 664.2