## Variances

It has long been noted that corpus frequencies, taken alone, give a very limited picture of a word's distribution in a corpus. As well as varying in raw frequency, words vary in the extent to which they are equally spread across the documents on the corpus. This 'burstiness' can be measured in a variety of ways (Church and Gale, "Poisson Mixtures", JNLE 1(2), 1996). One straightforward possibility is to take a large number of documents, all of the same length; count the frequency of a word in each of these documents; and calculate the (mean and) variance of this frequency.

The file presents the results of such an exercise. It is potentially of interest for various statistical approaches to text processing (e.g. as author identification and information retrieval) as well as for linguistic studies of how much semantic content different English words have.

#### Method

The first 5,000 words of all documents (=files) longer than 5,000 words in the written part of the BNC were taken. There were 2018 of these, so the subcorpus was slightly over 10M words. (I used written-only on the premise that the spoken material would be too different to usefully treat as part of the same population - of course, one might say this about all sorts of subcorpora, but never mind.) A frequency list was produced for each of these (truncated) documents. Then, taking the 8189 word-pos pairs occurring 100 times or more in the sample, a 2018x8189 table giving the frequency of each word in each document was produced. For each word, the mean and variance was calculated. There were two ways to calculate mean and variance: including the zeros (eg always dividing by 2018) or excluding them (dividing by the number of documents the word occurred in). For most purposes, it is the former that is of interest so this is what I present. The "exclusive" figures may readily be reconstructed.

**File format**

Columns are

(1) Word )Using BNC definitions of 'word' and tags

(2) POS-tag ) - see IJL paper (details above)

(3) Total freq (in 10M corpus)

(4) (Truncated) documents that word-pos pair occurs in (out of 2018)

(5) Mean (= Total freq./2018)

(6) Variance

(7) Variance/mean

The last is useful because, for distributions like the normal, poisson, binomial, variance increases with mean, so, to make the variance figures comparable for words of different base frequency, it is necessary to normalise by the mean. This is the figure that shows that, e.g., pronouns have very high variability, and prepositions, low (cf. Kucera and Francis 1982). Words are presented in frequency order. The file is 100 KB (compressed).