**Burrows Delta and how it relates to the clustering and the classification methods discussed in this and the previous chapters**

Documents are usually of unequal word length. Hence, when assessing the distance of word frequencies among two different documents, we recommend the use of document length-adjusted relative word frequencies as this standardizes the word frequencies for the length of documents. Several distance measures can then be applied to the vector of relative word frequencies such as the L2 (Euclidean) and the L1 (Manhattan) distance.

**Burrows delta**, a measure of closeness of two documents,does basically the same, but also adjusts the difference in the relative frequencies of each word by a measure of variability in the relative frequencies of that word across all documents. Stylometry tries to assess whether a document of unknown authorship is close or far from the known writings of a certain author. For example, how distant is the epic John Milton (1608-1674) poem "Paradise Lost" from his early writings? Yes, that poem was written by Milton, but let us assume here that we don't know. One can easily obtain all words in that poem and their relative frequencies (relative to the total number of words in that poem). Similarly, we can look at all early poems of John Milton, and determine its words and their relative (length-adjusted) frequencies. These two vectors of relative frequencies express the similarity of the two groups. One could apply one of many measures of distances such as the L2 (Euclidean) distance and the L1 (Manhattan) distance.

Burrows, in his original 2002 paper, adopts the L1 (Manhattan) distance, but includes one additional step. He makes use of a reference set of many Restoration (English 17th century) poems which provides him with a list of the 30 most frequently used words. For each word he calculates the mean and the standard deviation of the relative frequencies across all poems in this reference set. This provides a measure of variability across all poems in the reference group. He then calculates for each of the 30 most common words the standardized z-score of the relative word frequency in all early poems of John Milton (that is, he standardizes the difference between the frequency in Milton's early verse and the average relative frequency in the reference set by the standard deviation of the relative frequencies in the reference set). A z-score of +2, for example, tells us that the Milton frequency is 2 standard deviations larger than the average in the reference set). He calculates a similar z-score for the word frequency in the unknown poem "Paradise Lost". He does this for each word, and then averages the absolute values of the differences in the z-scores across the 30 words. The magnitude of that average expresses the closeness of the word structure of Milton's early verse and the poem "Paradise Lost". Doing so, Burrows incorporates the variation in the relative word frequencies across poems.

When we describe the distance between documents in Chapters 7 and 8 we could have used the whole corpus as a reference set. We could calculate for each word the standard deviation of its relative frequencies across all documents. We then could calculate a weighted average of the absolute differences in the relative frequencies of the words, with weights given by the inverse of their standard deviation. However, we don't expect the ordering of the distances to change by much. There could be small changes for very short documents that contain a very rare word as this modification would down-weight the difference between documents with and without that rare word.

Burrows J. 'Delta’: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing 2002; 17(3):267–287.