



Cluster Analysis

- **Cluster analysis of this paper**

- **Distance**

- **Euclidean distance:** appropriate for variables that are uncorrelated and have equal variances
- Statistical distance: adjusts for correlations and different variances

- **Linkage**

- Single linkage: based on the shortest distance between objects
- Complete linkage: based on the largest distance between objects
- Average linkage: based on the average distance between objects
- **Ward's method:** based on the sum of squares between the two clusters, summed over all variables
- Centroid method: based on the distance between cluster centroids

- **Standardized variables**

- **Hierarchical clustering**

- **Number of cluster: number of authors/styles**



Cluster Analysis

■ MINITAB

- All analyses in the paper were performed in MINITAB, including cluster analysis and PCA
- MINITAB is a general-purpose statistical computing system; little or no previous computer experience are required
- <http://www.minitab.com>



Burrows's Methods

- Aim to calculate “pure distance” or delta between two texts
- The results of the delta method seem quite promising with a high accuracy (~90%) even when comparing a number of authors.
- Use frequency information of the most commonly occurring function words



Burrows's Methods

■ Procedure

- Calculate frequencies of the most commonly occurring function words, both for a main-set of texts written by all of the authors in question and for the texts which are to be compared.
- Calculate the standard deviations for all of the words for the main-set. Then the z-score is calculated for each word for every piece of text. The z-score shows how many standard deviations from the mean a value is. z-score is calculated in the following equation:
 - $Z(w) = (\text{Freq}(w) - \text{mean}(w)) / \text{StandVar}(w)$
- Calculate delta score between two texts. A delta score is the difference between the z-scores for each word in the tested article and each author's work. Absolute values for these differences are calculated and the mean for all the words is produced.