ICHL 2007 Montréal
Plenary Sessions / Conférences plénières

Statistical Inference in Historical Text Mining

Ludovic Lebart
École Normale Supérieure des Télécommunications

Principal axes methods (multivariate descriptive techniques such as principal components analysis, correspondence analysis) provide useful visualizations of high-dimensional data. In the context of historical textual data, these techniques produce planar maps highlighting the associations between graphemes and texts (paragraphs, chapters, full texts, authors). The set of graphemes is frequently very large, and tools are needed to assess the locations of points and then select the subsets of units that are significant from a statistical standpoint. A classical analytical approach is both unrealistic and analytically complex. However, the "bootstrap" techniques as well as similar Monte-Carlo methods make weak assumptions about the underlying distributions and allow for drawing confidence areas for the locations of points in the obtained graphical displays. A valid statistical inference can then be carried out in a particularly complex context. Examples relate to a series of medieval French texts (12-th to 14-th centuries) rich in spelling variants. A free software is available.