

# Textual and lexical statistics

Mónica María Bécue Bertaut<sup>1,2,\*</sup>

1. Universitat Politècnica de Catalunya. Departament de Estadística i Inv. Operativa
2. IDT. Institute of Law and Technology of the Universitat Autònoma de Barcelona

\* Contact author: monica.becue@upc.edu

**Keywords:** textual data; textual statistics; correspondence analysis; lexicometry; constraint clustering methods

Statistics methods applied to the particular data that the texts are, in their very diverse forms, is a huge domain. In this work, we will focus on tools that belong to textual and lexical statistics.

The problems tackled are usually divided into two types: form versus content of texts. However, in fact, both aspects intertwine. A statistical approach is applied to such diverse sets of documents as classical works, political speeches, newspaper articles, collections of scientific research papers, closing speeches for the prosecutions in trials, free-answers to open-ended questions in surveys, short free-text comments in sensory data collection, etc. We can have to deal with a set of texts, or corpus, with objectives such as to detect similarities and differences, to build a partition of the texts into clusters and/or to characterize every text as compared to the others. Under other circumstances, we have to study a single text aiming at revealing its structure and evolution, that is, how the author has elaborated and organized the argumentation.

In every case, the searched information depends on the objectives and on the nature of the texts. This will drive the selection of the textual units (tool or/and full words; keeping all the words versus selecting particular words) and textual data preprocessing and coding.

Textual statistics adopt a multidimensional approach. The corpus to be analyzed is coded through a table documents×words. Correspondence analysis (Benzécri, 1976; Benzécri, 1981; Lebart & Salem, 1998; Murtagh, 2005), starting from the distribution of the different words in the texts or parts of the texts, is the key method in this approach. The present possibilities of the computers increase its potentiality to visualize the information extracted from the analyzed texts. Clustering, or constrained clustering, is usually associated to correspondence analysis to enrich and complete the interpretation.

Other methods, peculiar to the textual domain and grouped under the name of lexical statistics (Muller; 1977), are also profitable to extract information from the texts. Born around the project “*Trésor de la langue française*” (Treasure of the French language) in the fifty’s, these methods mainly study the richness, specificity, increase and evolution of vocabulary, that is, characteristics of the style of an author and adaptation to the circumstance of the audience and/or to the type of work.

Both groups of methods can be jointly used with profit. We will show the main results that they provide in the study of a closing speech on behalf of the prosecution in a lawsuit for murder. This speech has to prove a hypothesis, persuade and convince the audience. The strategy elaborated by the prosecutor leaves signs in the chosen words and their distribution within the text. To detect these signs allow for putting to the fore important rhetorical features. The whole of the methods help to reveal the evolution of the speech, locate the drawbacks and identify the moments of disruptions. This allows for segmenting the speech in homogeneous temporal periods that are, further, described by their characteristic words.

Other applications will be briefly mentioned to put to the fore the types of conclusions that can be drawn from statistical analyses of texts.

## References

Benzécri, J.P. (1976). *L’Analyse des Données II. Correspondances*, 2<sup>nd</sup> éd., Dunod. Paris.

Benzécri (1981). *Pratique de l’analyse des données. Tome 3. Linguistique & Lexicologie*. Dunod, Paris.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring textual data*, Kluwer, Dordrecht.

Muller, Ch. (1977). *Principes et méthodes de statistique lexicale*, Paris, Hachette.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall.