Processing texts and open-ended questions in sample surveys

Ludovic Lebart [1,*]

1. Telecom-Paristech, Paris, France
* Contact author: ludovic@lebart.org

Keywords: Texts, Open-ended questions, Visualization, Correspondence Analysis, Clustering

Open-ended questions can be found in the questionnaires of a multitude of surveys performed in socio-economy, epidemiology, advertising, marketing, sensometrics. They become an essential part of these questionnaires when a more fundamental research is involved and/or when a complex and unknown topic is being explored (particularly during periods of uncertainty, in the context of an unstable and risky environment). Processing "free responses" (i.e.: responses to open questions) is a challenge for both statisticians and specialists in text analysis.

When dealing with socio-economic sample surveys we are often facing the following situation: Numerous statistical observations are simultaneously described by texts and by several categorical variables. For instance, individuals in a survey are characterized by both responses to open-ended questions and responses to closed questions. In multinational or cross-cultural surveys, as well as in some industrial databases (complaints files, for example), it frequently happens that the texts are in different languages, although the categorical variables are the same.

The techniques that will be briefly presented take advantage of the presence of categorical variables to group responses and consequently to build as many artificial texts as there are categorical variables. The main task consists then in comparing texts, instead of understanding texts. Such purely comparative approach may allow for tackling multilingual surveys or corpora.

About the technical aspects of the methodology, let us mention that both principal axes techniques (factor analysis, principal component analysis, correspondence analysis) and clustering methods play a major role in the computerized exploration of textual corpora. They produce visualizations and groupings of elements (free responses in marketing and socioeconomic surveys); they highlight associations and patterns; they devise decision aids for attributing a text to a category of respondent or to a specific period.

We will focus our presentation on the use and assessments of visualization techniques in several practical situations. The examples of application concern texts derived either from food and wine guides or from open-ended questions in a series of international sample surveys.

References

Becue-Bertaut M, Alvarez-Esteban R, Pages J. (2008). Rating of products through scores and free-text assertions: Comparing and combining both. *Food Quality and Preference* **19**, 122–134.

Benzécri J.-P. (1992). *Correspondence Analysis Handbook*, Marcel Dekker, New York.

Greenacre M. (1993). *Correspondence Analysis in Practice*. Academic Press, London.

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

Lebart L., Morineau A., Warwick K. (1984). *Multivariate Descriptive Statistical Analysis,* Wiley, New York.