Local Course
Lexicometry and Discourse analysis

Barcelona, Wednesday 6th May – Thursday 7th May 2015

By Gaël Plumecocq – Fellow Researcher in Socio-Economics at the French Institute for Agricultural Research (INRA)

ABOUT THE COURSE

Lexicometry is the measurement of the frequency with which words occur in text. It provides statistical indicators and graphical representations that enables investigating written texts (novels or essays, legal texts, transcribed interviews, political discourses, scientific writings...): semantic classifications, factor analyses, similarity analyses, word clouds, lexical density, intertextual measures... These indicators and representations compress information contained in large texts and provide new perspectives to analyze them.

Lexicometry methods also allow introducing contextual variables such as, e.g the socioeconomic characteristics of the speakers (age, sex, income, belonging institution, hair color...). Textual analyses then put face to face the meaning of words used in texts (semantic dimension) with the socioeconomic context of their enunciation/production (institutional dimension). In that sense, lexicometry not only provides quantitative measurements and indicators of texts, it also provides basis for qualitative analysis. It may then be used either to formulate theoretical propositions in an inductive way or to explore hypotheses via the introduction of variables.

The tools presented in this course enables analyzing every kind of text written in latin, cyrillic, or Greek alphabets. Texts written in any kind of language can be analyzed as long as they are transcribed in those alphabets (except for agglutinative language such as Basque, Finnish, Turkish or Japonse). Texts must be in .doc, .html, or .txt formats (.pdf sometime work but introduces artifacts in texts).

Gaël Plumecocq is fellow researcher in economics at the French Institute for Agricultural Researches (INRA). He is involved in 4 research projects funded by the French Agency (ANR), and a research project funded by the FP7 Program. In these works, he analyzes the ways in which collective judgments are formed (on sustainable performance or on products quality), involving knowledge and moral norms, as a condition of possible collective actions. To formally represent this issue, he uses lexicometric methodology, multicriteria assessment, and integrated evaluation frameworks.

This activity has received funding from the Seventh Framework Programme of the European Commission (Marie Curie Actions). Contract n. 289374 (ENTITLE)

AGENDA

Location: Room Z/033, ICTA-ICP Building, Campus UAB, Bellaterra (Barcelona)

Wednesday 6 May 10h – 13h:

10h – 11h30: Theoretical foundations and conceptual implications of text analysis.

We will examine the various types of text analysis, as well as the underlying methodologies and theories (discourse vs. content analysis), the usefulness of lexicometry as regard to the multiple dimensions of the process of meaning production, and the conceptual problems arising when we try to measure vocabulary (tokenization, lemmatization...).

11h30 – 13h: Learning how to use Iramuteq, (toolbox for textual statistic):

In this part, we learn how to format texts to be analyzed, how to introduce coding variables, the various functionalities of the toolbox, the indicators and representation it provides, and how to use parameters to get the best of the information contained in texts.

Wednesday 6 May 14h30 – 18h:

14h30 – 17h: Statistical indicators and graphical representations (overview)

This part of the course presents the main indicators and visual representations provided by Iramuteq, their limitations and the conditions under which they provide useful information. It provides the statistical backgrounds to understand how they are constructed.

- Descriptive statistics: counting of occurrences of token-words, and grammatical categories.
- Alceste classification: descending classification of 'co-texts' assessed through chi-square test.
- Factor analysis: analysis of the variability of discourses and variables according to explicative factors.
- Similarity analysis: graph representing the relationship between words.

17h-18h: Questionnaire analysis with Iramuteq

Iramuteq can also process questionnaires or multivariate data (including answers to open questions). We briefly overview the possibilities of analyzing such data.

Thursday 7 May 10h – 13h:

10h- 13h: Learning how to use Iramuteq, (toolbox for textual statistic):

The whole morning will be dedicated to explore the functionalities of Iramuteq (http://www.iramuteq.org/), a lexicometry freeware interface for R. Participants are welcome to use their own textual corpuses to explore the various tools provided by Iramuteq. It provides tool to decompose and analyze the structure of texts in terms of semantic content (which are the various dimensions of a given text?), but also in terms of quantified repartition (which dimension is the most important, and which ones are less present in texts?) and in socioeconomic terms (which variable is the most representative of which semantic dimension?).

This activity has received funding from the Seventh Framework Programme of the European Commission (Marie Curie Actions). Contract n. 289374 (ENTITLE)