INTRODUCTION TO TEXTUAL DATA ANALYSIS WITH TEXTOMETRY

Àâòîð Ñåðæ Ýéäåí (Serge Heiden)
25.10.2015 ã.
Ïîñëåäíåå îáíîâëåíèå 28.10.2015 ã.

Textometry has been developed in France since the 70's through pioneering works in statistical analysis of lexical data (P. Guiraud & C. Muller) and general data analysis (J-P Benzécri).

Textometry combines several computerized tools to assist the systematic analysis of large bodies of digitized texts in any language. Texts are organized in sets called corpora. They can come from written language productions (books, news, letters...) or from transcriptions of speech (political discourse, interview...). They are digitally encoded according to some conventions like XML encoding according to the TEI guidelines that can be used in the analysis. They contain words that are used to access their content and style. Words are automatically segmented and tagged with properties like a grammatical category or a lemma. Documentary tools use any combination of those word properties to list their patterns and display in which contexts they occur (frequency lists, concordance and edition display). Statistical tools apply statistical models to the counting of word properties to analyze their distribution across corpora (factorial analysis, clustering), their over- or under-representation in some sub-corpus (specificity analysis) or to analyze attraction between word patterns (cooccurrence analysis).

All statistical tools results are related to documentary tools to give access to the detail of any textual event. As a result, textometry provides a comprehensive set of highly synthetic or very precise views of text contents.

TXM implements the textometry tools as an end user desktop software for Windows, Mac or Linux. Its Graphical User Interface provides common services for interacting with the software: menus to launch commands, views to display corpora, text editions and tools results. At the heart of TXM, the CQP search engine helps the user to list any word patterns occurrences in a corpus to list, count and display them. The workshop introduces sequentially to the following concepts and tools:

- whole lexicon and frequencies display;

- index and frequencies of some patterns expressed by CQL expressions for the search engine;

- concordance display of CQL patterns;

- reading of text editions;

- progression analysis of some CQL patterns;

- sub-corpus building;

- specificity analysis;

- cooccurrence analysis.

Each command result, as a table of numbers or as a graphic, can be exported to be analyzed or edited by another tool.

TXM version as a web portal software - to allow on line access - is finally introduced.

To be able to be analyzed by TXM, a corpus must first be imported into it.

TXM can import and analyze two main types of textual data sources:

- written texts: which can be the result of an OCR process with the associated images of the original texts, and can have various levels of encoding (from raw text as a base stream of characters to XML encoded representations according to the TEI guidelines)

- speech transcriptions: in which the speech turns of each participant is delimited and can be synchronized to the original audio or video record.

The result of any import process into TXM is a new corpus containing the elements of its corpus data model:

- a corpus contains several texts that can each have several properties (called metadata), such as an author, date, title, genre, etc.

- a text can contain several internal structures that can each have several properties, such as a chapter, section, paragraph, speech turn, etc.

- a text contains words that can have several properties, such as the graphical form, the grammatical category and lemma

The workshop introduces progressively to the following concepts and tools:

- simple import from the system clipboard (text copied from any application: word processor, browser, mail client, etc.);

- import of TXT raw text files and tuning the language for automatic lemmatization of words;

- import of TXT raw text files and tuning texts metadata;

- import of XML files and tuning internal structures;

- import of XML files and tuning some specific words pre-encoding.

For each level of sources imported, the different elements of the corpus data model available in TXM are verified and manipulated by tools to show the trade off between the work of source texts encoding and what is necessary for the analysis of a particular corpus using TXM

Ïðîãðàììà êóðñà

«Introduction to textual data analysis with textometry»

(6 ÷àñîâ ëåêöèîííûõ çàíÿòèé è ïðàêòèêóìîâ, ñðîêè: 19-21 íîÿáðÿ 2015 ã.)

Ëåêöèÿ 1: Introduction to textual data analysis with textometry.

Ïðàêòèêóì 1: TXM software initiation hands-on workshop

Ïðàêòèêóì 2: Textual corpora preparation & importation into TXM software hands on workshop

Ëèòåðàòóðà

1. Heiden, S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In K. I. Ryo Otoguro (Ed.), 24th Pacific Asia Conference on Language, Information and Computation - PACLIC24 (p. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan. <http://halshs.archives-ouvertes.fr/halshs-00549764/en>

2. http://textometrie.ens-lyon.fr/spip.php?article9&lang=en (NB!)