**HEIDEN — TXM**

<0 images>

**Introduction to the TXM Content Analysis Software**

Heiden, Serge

The objective of the Introduction to the TXM Content Analysis Software tutorial is to introduce the participants to the methodology of textometric content analysis (<http://textometrie.ens-lyon.fr/?lang=en>) through the use of the free and open-source TXM software (<http://sourceforge.net/projects/txm/files/documentation/TXM%20Leaftlet%20EN.pdf/download>) directly on their own laptop computers. At the end of the tutorial, the participants will be able to input their own textual corpora (Unicode-encoded raw texts or XML tagged texts) into TXM and to analyze them with the panel of content analysis tools available: word patterns frequency lists, kwic concordances and text browsing, rich full text search engine syntax (allowing to express various sequences of word forms, part of speech, and lemma combinations constrained by XML structures), statistically specific sub-corpus vocabulary analysis, statistical collocation analysis, etc.).

During the tutorial, each participant will use TXM (from <http://sourceforge.net/projects/txm>) and the TreeTagger lemmatizer (<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger>) on her Windows, Mac, or Linux laptop and will leave the tutorial with a ready-to-use environment.

The tutorial will also introduce the participants to the TXM community ecosystem (users mailing list and wiki, bug reports, etc.) and to the TXM portal version server software (see, for example, <http://portal.textometrie.org/demo>) for online corpus distribution and analysis. Time permitting, TEI encoding aspects of corpora related to TXM could also be introduced, as well as speech transcriptions or parallel corpora encoding and analysis.

Such tutorials have been given on a monthly basis in Lyon (France) since September 2012 (see, in French, <https://groupes.renater.fr/wiki/txm-users/public/ateliers_txm>).

It has proven to be very beneficial to participants from various fields of the humanities working on digital textual data: geography, history, linguistics, literary studies, sociology, psychology, urbanism, political sciences, economy, etc.

The tutorial will be taught in English and will complement two accepted communications introducing the TXM platform given during the conference:

• ‘Progressive Philology with TXM: From “Raw Text” to “TEI Encoded Text” Analysis and Mining’, #463.

• ‘From KWIC Concordance to Video Excerpt or Folio Facsimile: Demonstration of Multimodal and Multimedia Corpora in TXM’, #468.

**Tutorial Instructor**

Serge Heiden (slh@ens-lyon.fr)

Project Manager of TXM Platform Development (<http://textometrie.ens-lyon.fr/spip.php?article9>)

*S. Heiden* develops the textometry content analysis methodology in a research team of five people through the development of tools able to process richly encoded corpora (<http://icar.univ-lyon2.fr/pages/equipe31.htm>). Working on the relation between analysis tools and XML-TEI encoded corpora, he is involved in the TEI consortium activities as the TEI Tools SIG convener (<http://www.tei-c.org/Activities/SIG/Tools>).

**Target Audience and Expected Number of Participants**

Participants can come from any humanities and social sciences disciplines. No previous statistical or XML background is necessary. Participants can come with their own corpora.

The ideal number of participants is about 12 to 15 people; the maximum number of participants is about 20.

Each participant should come with her own laptop computer.

The tutorial needs to run at least for a full day\*: typically half day for TXM tools fundamentals and half day for main corpus formats fundamentals (TXT and XML) and input procedures into the platform.

\*The regular TXM tutorials run for two days (one-day TXM introduction, one-day corpus formating and import into TXM).

**Brief Outline**

*9am–12pm (3h) + 1pm–5pm (4h) = 7h total*

*Install and introduction: 45 minutes*

• TXM, TreeTagger, sample corpus installation checkup (participants will be asked to install the software before coming to the workshop to save time).

• TXM user interface & windows, corpus Description command.

*Main tools: 2 hours, 15 minutes*

• Lexicon analysis and spreadsheet export.

• Index building for distributional semantics and Corpus Query Language syntax.

• Concordance and Edition browsing, Progression graphics.

• Sub-corpus building, corpus partitioning, and specificity/factorial analysis/clustering.

• Words co-occurrence analysis.

• TXM portal demo (optional).

• TXM community: mailing lists, websites, and documentation.

*Importing corpora into TXM: 4 hours*

• TXM import strategy and main corpus formats: TXT-Unicode+CSV, XML+CSV, XML-TEI: 1/2 hour.

• TXT-Unicode sample corpus and TXT+CSV import into TXM, sample analysis: 1 hour, 15 minutes.

• Introduction to XML and to TXT2XML conversion tools: 1/2 hour.

• XML sample corpus and XML/w+CSV import into TXM, sample analysis: 1 hour, 45 minutes.