

# Exploiting TEI-annotated data with TXM

Serge Heiden<sup>1</sup>

September 2013

CC-BY

TXM is a free and open-source textual data analysis platform compatible with TEI encoded sources (<http://sourceforge.net/projects/txm>). It has already been applied to a dozen different applications of the TEI guidelines :

- Perseus: <http://www.perseus.tufts.edu/hopper>
- TextGrid: <http://www.textgrid.de/en>
- NLTK - Brown Corpus (TEI XML Version):  
[http://nltk.googlecode.com/svn/trunk/nltk\\_data/index.xml](http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml)
- Frantext (libre): <http://www.cnrtl.fr/corpus/frantext>
- Base de Français Médiéval (BFM): <http://bfm.ens-lyon.fr>
- BVH Epistemon: <http://www.bvh.univ-tours.fr/Epistemon>
- Bouvard&Pécuchet: <http://dossiers-flaubert.ish-lyon.cnrs.fr>
- Presses Universitaires de Caen (PUC), MRSH de Caen - Revues.org:  
[http://www.unicaen.fr/recherche/mrsh/document\\_numerique/outils](http://www.unicaen.fr/recherche/mrsh/document_numerique/outils) ([DISCOURS scientific journal])

## What is Querying and what is it for?

Research projects using TXM on their TEI-annotated corpora apply various tools implementing what we call the 'textometry' methodology. The idea is to mix qualitative tools like concordances or word lists and quantitative tools like statistical cooccurrences analysis or statistically specific words of a sub-corpus analysis. In this framework, querying data is at the core of the methodology and is used for various services:

- query for **display**: for example word contexts synthesis (*concordances*) or *text edition* pages display with matching words highlighted;
- query for **counting**: for example word *frequency lists*, various *statistics* based on word frequency.

## Querying in What?

Applying the 'textometry' methodology supposes to build and use various *sub-corpora* and *partitions* (contrasts between texts or parts of texts) depending on the data and the kind of analysis. A partition can be seen as a set of sub-corpora the sum of which is the whole corpus. Each sub-corpus forms a set into which querying are performed and the sub-corpus semantics is as important to be defined and managed as the queries that can be done in it.

## What Relations between TEI-annotated data and query technologies?

TEI-annotated data have three main interlinked semantic levels:

---

<sup>1</sup> I would like to thank Alexey Lavrentiev for fruitful comments on drafts of this proposal.

- Unicode text level semantics;
- XML text level semantics;
- TEI text level semantics.

Querying technologies can address data at any of those levels and combine them for different purposes. In TXM, instead of building a universal querying system, the strategy is to use specific technologies<sup>2</sup> for specific uses and to relate data sources to them through different paths. For the moment, Unicode and XML semantic levels are not used directly, as respectively Lucene and XQuery search engines would do. A complementary data format is also used for syntactically annotated data (TIGER-XML):

- for *annotated word sequence* queries, we have put the CQP search engine at its core. This search engine was chosen for its best ratio of expressiveness of queries / efficiency of resolution in space and time. All base tools of the 'textometry' methodology rely on queries and results offered by this search engine (word lists, concordances, lexical tables, etc.). This allows the user to easily built the table of verb lemma frequencies between some texts for example. As CQP manages one simple layered hierarchy, the user can express queries with the help of some structures inside texts. Structures are also queried by CQP to build various corpus configurations into TXM: *sub-corpora* and *partitions*. The relation between TEI-annotated data and the CQP search engine is established through the TEI extension format called "XML-TXM". This format, developed and maintained by us, unambiguously encodes text structures and words. Word level annotations use a standoff encoding scheme for ease of maintenance between NLP tools application on the sources;
- for *aligned corpora*, we use the same CQP search engine which allows us to express queries between two aligned corpora. The XML-TXM format also uses a standoff scheme for alignment between structures of texts in parallel corpora;
- for *syntactically annotated* data queries, we have chosen the TIGERSearch engine as a complementary one to CQP. For that search engine the source data format is the TIGER-XML format and the relation between the TEI / XML-TXM sources and the TIGER-XML sources is established at the word level: words share the same XML identifiers so that the two search engines can be used complementary on the same corpora but for different purposes and through different tools.

The querying system built on those two search engines is completely integrated for the user through the TXM Graphical User Interface. Queries to the CQP search engine are done through a graphical query builder assistant or directly with Corpus Query Language (CQL) expressions input. The raw output of CQP queries consists of a list of word tokens intervals and is fed to the various tools for the end user (HTML text editions browsing with matches highlighted, Concordances to display surrounding context, tabulated word patterns lists, sub-corpus building, partition building, etc.). Queries to the TIGERSearch engine are done independently with TIGERSearch expressions input. The raw output consists of a list of matches in syntactic trees and is fed to the various tools for the end user (matching trees display and browsing, "syntactic" Concordances to display surrounding context and sub-matches). The query expressions of the two search engines and the respective annotation models are currently not combined. We plan to add a complementary XQuery search engine to give access to the intermediary XML-TXM representation of corpora (compatible with the other search engines corpus model) and eventually to the TEI sources themselves (but with no direct compatibility with the other search engines).

## **Discussion**

In TXM data flow architecture information moves around from data sources to technologies and

<sup>2</sup> Being open-source, we also try to delegate the development and the maintenance of some components to efficient communities.

annotations are selected, adapted and filtered for different purposes. Each querying technology has a focus on the data sources and not every annotation is pertinent or compatible for all querying systems. At any given moment, a TEI annotation may be useless or not pertinent from one usage perspective and useful from another. In a way, this is similar to what happens to punctuations in medieval texts queried for syntactic analysis. TXM users asked to get rid of the linguistic “annotation device” called “punctuation” (comma, points...) added by scientific editors to manuscript editions to ease syntactic annotations querying, while asking for punctuations to come back in concordances to speed up contexts reading.