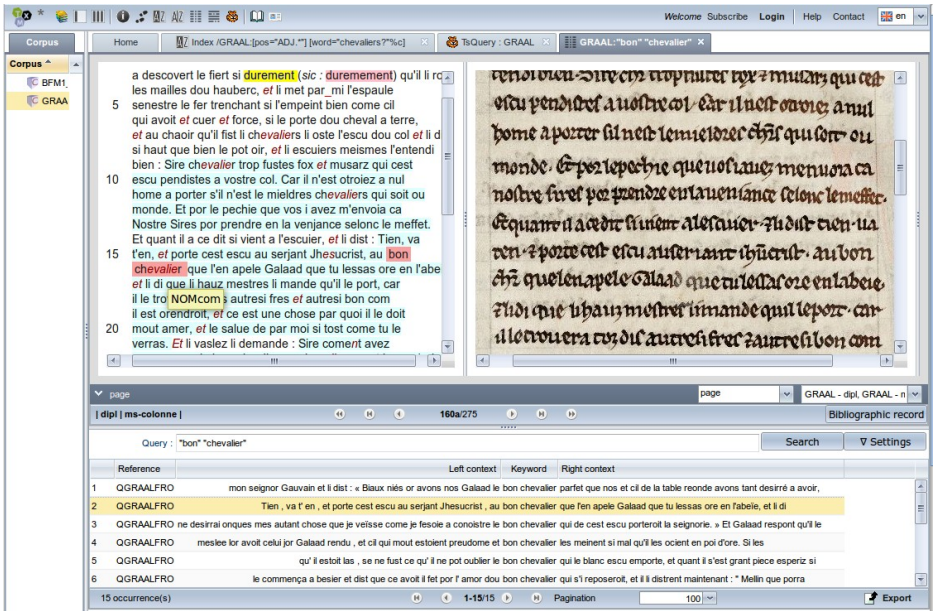
The TXM platform is a Unicode - XML & TEI compatible text/corpus analysis environment and graphical client based on the CQP search engine and the R statistical environment.

Supported platforms

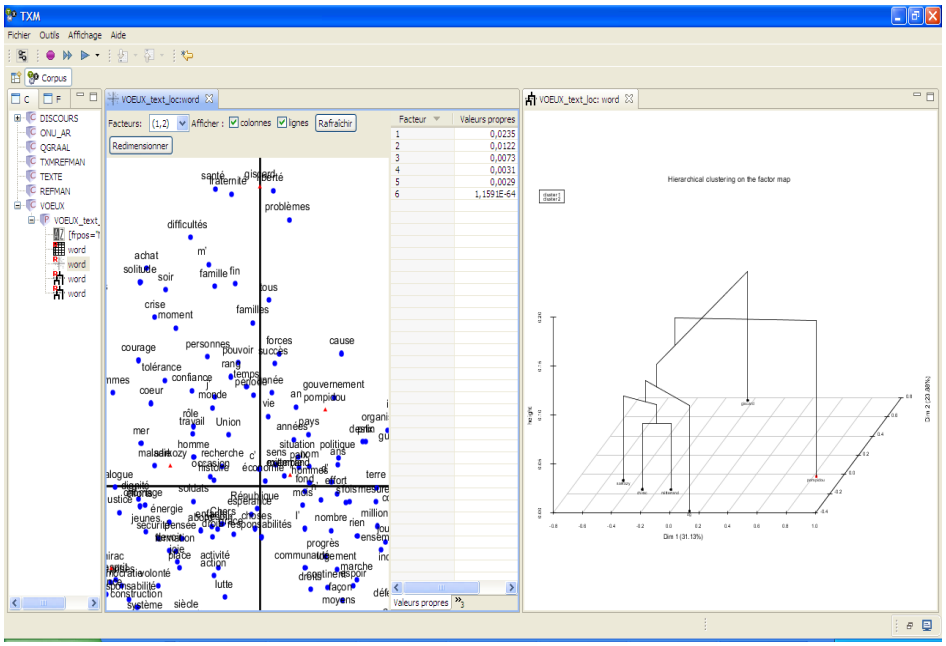
The standalone version runs on:

- **Windows** - 32bit or 64bit (tested on XP, Vista and 7)
- **Mac OS X** (tested on 10.5, 10.6 and 10.7)
- **Linux** - 32bit or 64bit (tested on Ubuntu and Debian)

The portal server runs on any J2EE conformant platform (tested in **Tomcat** and Glassfish on Linux and Windows).



XML-TEI Critical edition and Concordances in TXM web portal



Factorial plane and Clusters dendrogram

User Interface Languages

- English (EN)
- French (FR)
- Russian (RU)

How to download

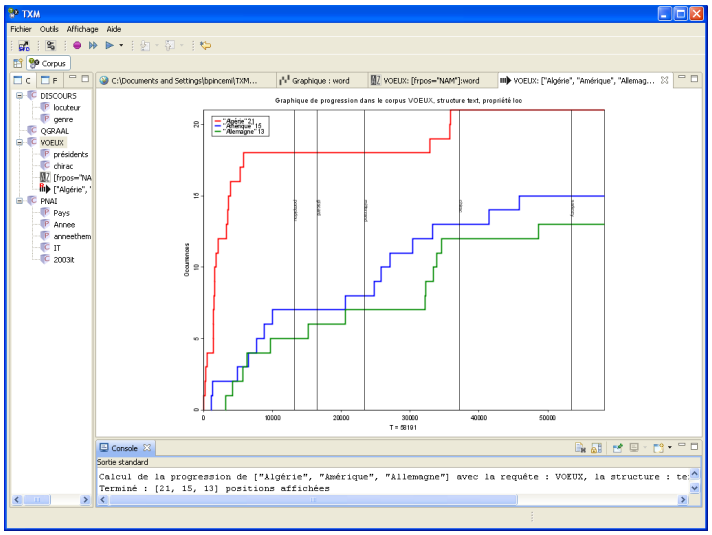
TXM is free to download and use, it is distributed as **open-source** software under the GPL V3 license: http://sourceforge.net/projects/txm

Contact

- Research project web site: http://textometrie.ens-lyon.fr/?lang=en
- TXM software web site: https://sourceforge.net/projects/txm
- Contact email: textometrie@ens-lyon.fr

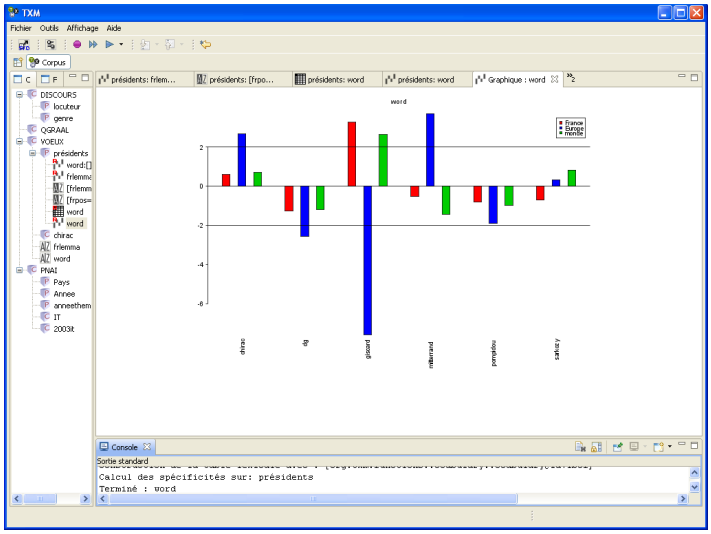
Features

- Provides QUALITATIVE ANALYSIS tools:
  - kwic **concordances** of word patterns based on the efficient CQP full text search engine and its powerful CQL query language
  - word pattern **frequency lists** based on any word property (graphical form or type, lemma, pos...)
  - word pattern **progression graphics**
  - examples of word patterns, expressed in the CQL query language which is based on word & structural level properties:
    - `"aiming"` to simply search for the word 'aiming'
    - `".*ing"` to search for words ending in "ing" (including mainly verb forms)
    - `[pos="VERB" & word=".*ing"]` to search for verb forms ending in ".ing" (where Part of Speech annotation is present)
    - `[lemma="group"] []{0,3} [pos="VERB" & word=".*ing"]` to search for the collocation <group lemma> followed by a <verb with progressive aspect> with at most 3 words in between
  - rich HTML-based text edition navigation with links from all other tools



Progression graphics

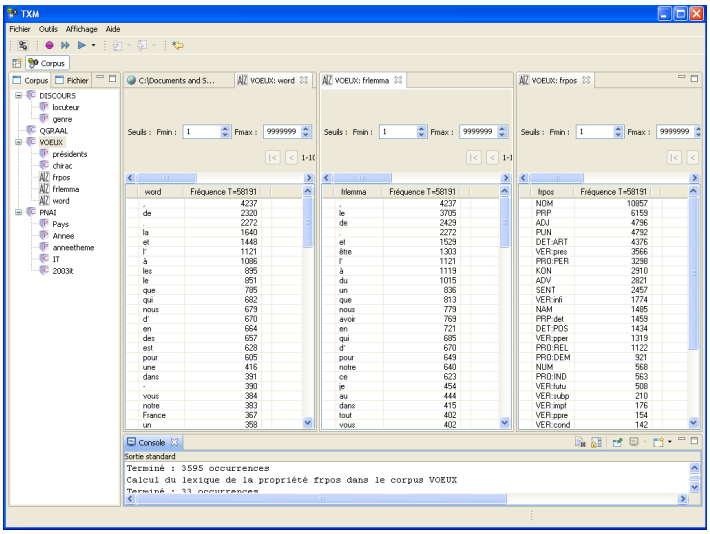
- Provides QUANTITATIVE ANALYSIS tools, based on R packages:
  - **factorial correspondence analysis**
  - **cluster analysis**
  - **specific** word patterns analysis
  - **collocations** analysis



Words Specificity diagram

- Helps to build various CORPUS CONFIGURATIONS: **sub-corpora** or **partitions** (for contrastive analysis between text structures or word selections)
- **Export**s any result in CSV, XML or SVG format
- Provides a large spectrum of INPUT FORMATS
  - several text formats (from raw to rich):
    - **Unicode TXT**
    - **ODT**
    - **XML**
    - **XML/w** (where some or all word limits and properties can be pre-encoded)
    - XML-**TEI P4** (according to Perseus project practice)
    - XML-**TEI P5** (according to various projects practice: BFM, BVH, NLTK, etc.)

    - speech transcription: XML-**TRS** (from Transcriber software, with time synchro)
    - aligned corpora: XML-**TMX** (with texts in relation of translation or versioning)
    - news portal articles: XML-**PPS** (Factiva), Europresse
    - etc.
- Applies various NLP TOOLS on the fly on texts before analysis (e.g. **TreeTagger** for lemmatization and pos tagging)
- Provides a RICH DATA MODEL: words and their properties inside hierarchical structures of texts with external or internal text metadata or speaker metadata
- Provides SCRIPTING facilities for repetitive or lengthy tasks automation or for platform extension (in **Groovy**/Java dynamic language)
- Includes a complete **text editor** to edit data sources, results and scripts



Lexicons compared side by side

Support for XML - TEI

Supports various flavors of TEI P4/P5 encoding practices:

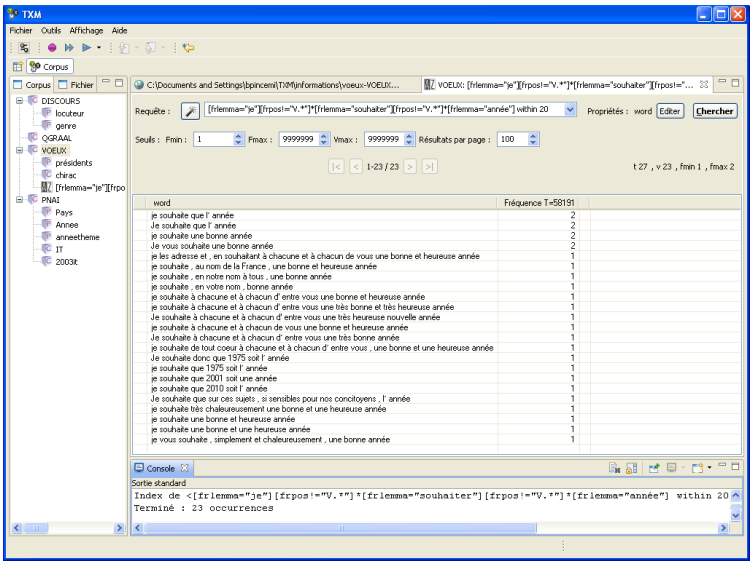
- Perseus: http://www.perseus.tufts.edu/hopper
- TextGrid: http://www.textgrid.de/en
- NLTK - Brown Corpus (TEI XML Version): http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml
- Frantext (libre): http://www.cnrtl.fr/corpus/frantext
- Base de Français Médiéval (BFM): http://bfm.ens-lyon.fr
- BVH 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 ([DISCOURS journal])
- TXM: https://sourceforge.net/apps/mediawiki/txm/index.php?title=XML-TXM

TEI sources are preprocessed by several XSL stylesheets, one can find in TXM source code. Some of those stylesheets are available in the online TXM XSL stylesheets library: http://sourceforge.net/projects/txm/files/library/xsl

Text/Corpus Languages

TXM works natively with any Unicode-conformant corpus, including Right to left writing systems.

Language support at word level is specific to each NLP tool used (for example, TreeTagger can tag the following languages: BG, DE, EN, ES, ET, FR, FRO, GL, IT, LA, PT, RU, SW, ZH).



Word sequence patterns frequency list

Documentation

- Main entry point for documentation on TXM at the Textométrie project web site: http://textometrie.ens-lyon.fr/spip.php?article98&lang=en
    - See for example the TXM manual (in French) at http://txm.svn.sourceforge.net/viewvc/txm/trunk/doc/Manuel%20de%20TXM%200.7%20FR.pdf?revision=2332
- TXM user community wiki (in French) at https://listes.cru.fr/wiki/txm-users (includes a FAQ)
- TXM developers wiki (in English) on Sourceforge : http://sourceforge.net/apps/mediawiki/txm
- All available documentation (for users and for developers) published on Sourceforge: http://sourceforge.net/projects/txm/files/documentation

Tech support

Tech support is mainly provided through two mailing lists (see below).

Public access to Bug reports and Feature requests:
https://forge.cbp.ens-lyon.fr/redmine/projects/txm/issues

User community

Currently, the TXM user community communicates using two mailing lists and a wiki:

- International *mailing list* : txm-open AT lists.sourceforge.net (low activity)
    - See archives at http://sourceforge.net/mailarchive/forum.php?forum_name=txm-open
- The mostly French-speaking **mailing list** : txm-users AT cru.fr (most active)
    - See archives at https://listes.cru.fr/sympa/arc/txm-users
- TXM user community **wiki** (in French) at https://listes.cru.fr/wiki/txm-users

Training in the use of TXM is available every month in dedicated workshops in our laboratory, or every year at the CNRS summer school « Computing and Statistical Methods in Text Analysis » (MISAT), see http://laseldi.univ-fcomte.fr/ecole.

Current version number and date of release

- Standalone: 0.7.2 released on Tuesday 2nd July 2013
- Portal: 0.4 released November 2011

Grants

- Jan 2007- Dec 2011: S. Heiden, Textométrie research project - TXM platform development kickoff, French National Research Agency (ANR) grant #ANR-06-CORP-029;
- Jan 2012 – Dec 2014: D. Peschanski, Matrice research infrastructure - TXM platform development for historians, ANR grant #ANR-10-EQPX-21-01.