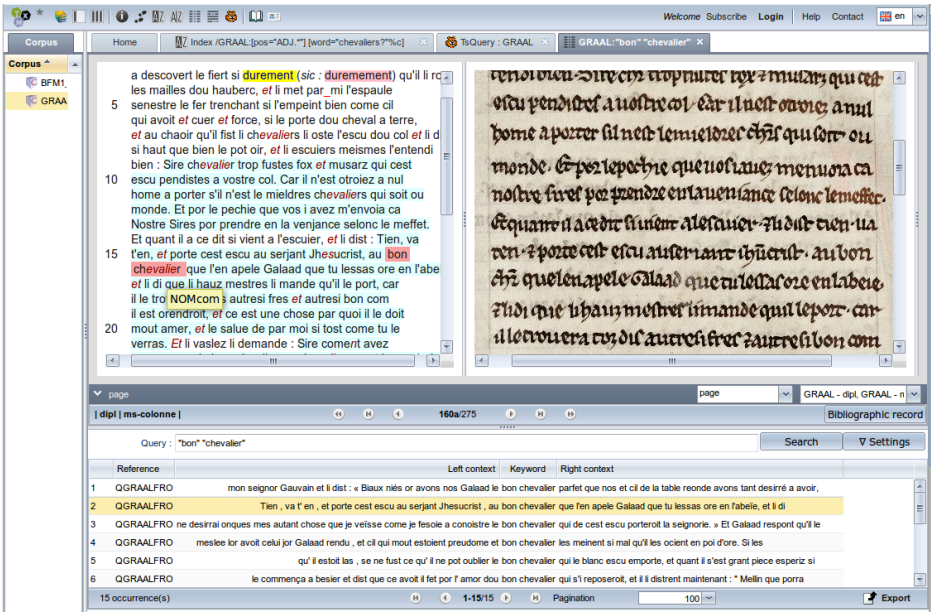
Logiciel d'analyse et de mise en ligne de corpus textuels

compatible Unicode – XML & TEI,
basé sur le moteur de recherche plein texte CQP
et l'environnement statistique R.

Le logiciel TXM a été conçu pour reprendre la tradition lexicométrique (implémentée notamment par Hyperbase ou Lexico 3) dans un contexte nouveau :

- corpus enrichis et structurés ;
- développement ouvert et communautaire.

Il a été initié dans le cadre du projet ANR Textométrie (2007-2010) et continue son développement grâce à son réseau de partenaires et notamment au soutien de l'Équipex Matrice (2012-2014).



Édition critique XML-TEI et Concordances dans la version portail

Systèmes supportés

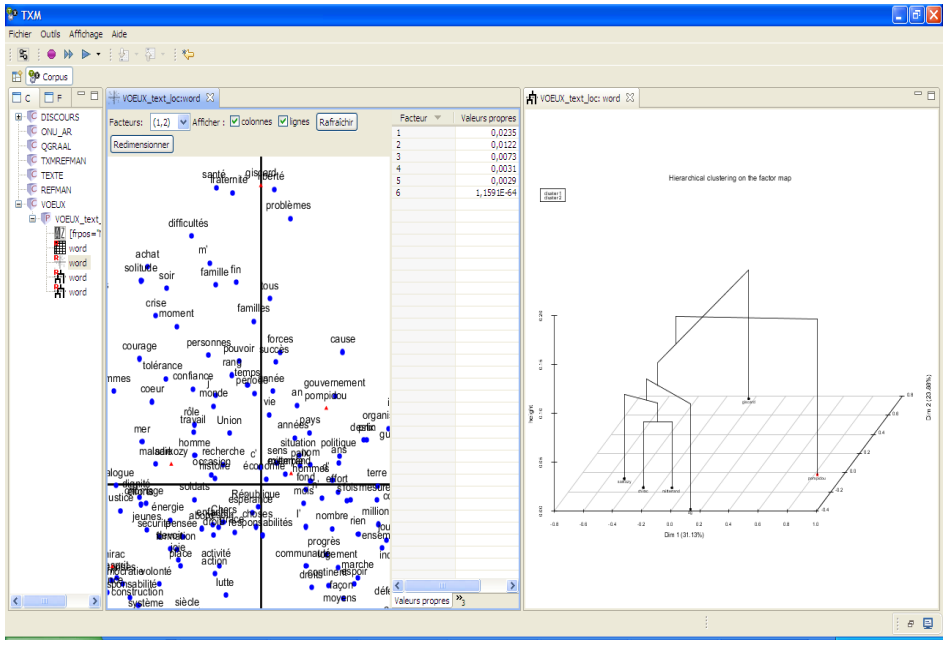
Version pour *poste de travail* :

- **Windows** - 32bit et 64bit (XP, Vista, 7 et 8)
- **Mac OS X** (10.5, 10.6, 10.7 et 10.8)
- **Linux** - 32bit et 64bit (Ubuntu et Debian)

Version *portail web* disponible pour serveur J2EE (Tomcat ou Glassfish sous Linux et Windows).

Langues de l'Interface

- Anglais (EN)
- Français (FR)
- Russe (RU)



Plan factoriel et dendrogramme de classication

Où le télécharger ?

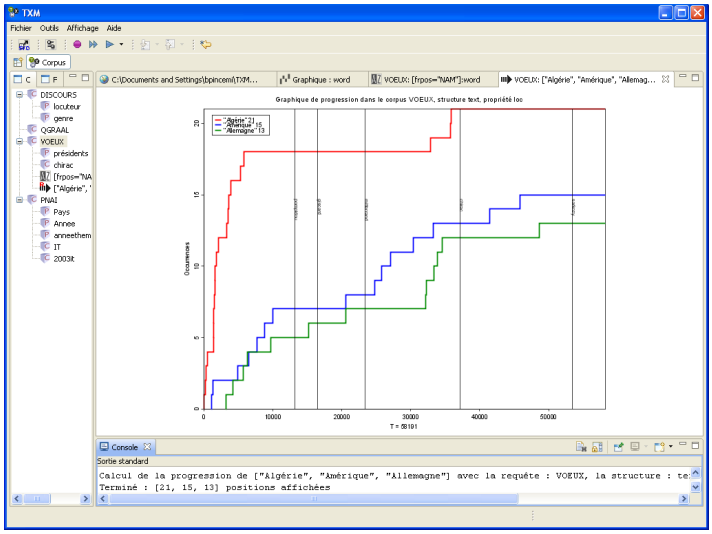
TXM est un logiciel *gratuit* diffusé sous licence *open-source* (GPL V3) à l'adresse : http://sourceforge.net/projects/txm

Contact

- Site du projet scientifique : http://textometrie.ens-lyon.fr
- Site du logiciel TXM : https://sourceforge.net/projects/txm
- Contacter l'équipe: textometrie@ens-lyon.fr

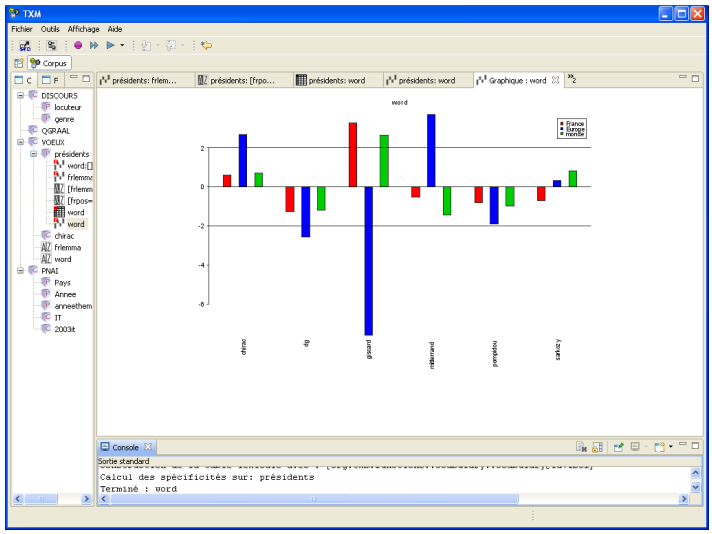
Caractéristiques & Outils proposés

- Outils d'ANALYSE QUALITATIVE :
  - **Concordances** KWIC de patrons de recherche de mots, à l'aide du moteur de recherche plein texte CQP et de son langage de requête CQL.
  - **Listes de fréquence** de patrons de recherche de mots combinant différentes propriétés de mots (forme graphique, lemme, catégorie grammaticale...)
  - **Graphique de progression** de patrons de recherche de mots (voir figure ci-contre)
  - Exemples de **patrons de recherche de mots** exprimés avec le langage de requête CQL (basés sur des propriétés de mots et de structures textuelles) :
    - `"souhaiter"` pour chercher simplement le mot « souhaiter »
    - `"souhait.*"` pour chercher les mots qui commencent par « souhait »
    - `[frpos="V.*" & word="souhait.*"]` pour chercher les verbes à n'importe quel temps de conjugaison qui commencent par « souhait » (pos = Part of Speech)
    - `[frlemma="je"] []{0,3} [frlemma="souhaiter"]` pour chercher le mot « je » suivi du verbe « souhaiter » à une distance de 0 à 3 mots
  - **Navigation hypertextuelle** au sein d'éditions de texte riches en HTML et liens vers différents outils de TXM.



Graphique de progression d'usage de mots

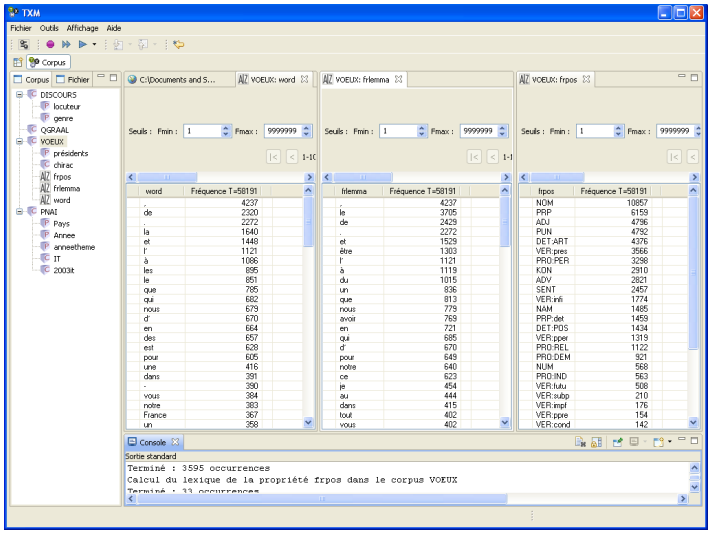
- Outils d'ANALYSE QUANTITATIVE basés sur des packages R(http://www.r-project.org)
  - **Spécificité** de patrons de recherche de mots
  - Analyse de **Cooccurrents** de patrons de mots
  - **Analyse Factorielle des Correspondances (AFC)**
  - **Classification Ascendante Hiérarchique (CAH)**



Histogramme de spécificité de mots par parties

- Création à la demande de CONFIGURATIONS DE CORPUS : **Sous-corpus** ou **Partitions** (pour les calculs contrastifs entre textes, structures textuelles ou sélections de mots)
- **Export** de tous les résultats de calculs aux formats tableaux (CSV) ou graphiques (SVG, JPG).
- Large spectre de FORMATS DE SOURCES pris en charge :
  - Corpus de textes (du plus élémentaire au plus riche) :
    - **TXT** avec encodage des caractères **Unicode** (texte brut)
    - **XML** tout venant (avec possibilité de pré-codage de certains mots avec une balise <w>)
    - XML-**TEI P4** (compatible avec les pratiques d'encodage du projet Perseus)
    - XML-**TEI P5** (compatible avec les pratiques d'encodage des projets BFM, BVH, TextGrid, NLTK, etc.)

    - Corpus de transcriptions d'audio ou de vidéo:
        - XML-**TRS** (du logiciel Transcriber, avec synchronisation temporelle)
        - TXT/ODT/RTF (convention de transcription en texte brut)
    - Corpus alignés : XML-**TMX** (textes en relation de traduction ou de version alignés au niveau des paragraphes, phrases...)
    - Corpus d'articles de presse issus de portails en ligne : XML-**PPS** (Factiva), Europresse
    - etc.
- Application automatique de différents outils de TAL sur les textes lors de l'import de corpus (e.g. **TreeTagger** pour la lemmatisation et l'étiquetage morphosyntaxique)



Comparaison côte à côte de lexiques

- Modèle de traitement de corpus riche : métadonnées de textes, propriétés de structures textuelles et propriétés de mots
- Pilotage automatique de la plateforme par **scripts** Groovy ou R (pour les sessions de travail répétitives ou longues, ou pour l'extension de la plate-forme)
- **Éditeur de texte** intégré : pour éditer les sources de corpus, les résultats de calculs ou les scripts

Compatible XML - TEI

Spécifiquement conçu pour être compatible avec des sources richement encodées en XML – TEI, TXM supporte différentes pratiques d'encodage TEI (P4 ou P5) :

- 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])
- Frantext (libre): http://www.cnrtl.fr/corpus/frantext
- Perseus: http://www.perseus.tufts.edu/hopper
- TextGrid: http://www.textgrid.de/en
- NLTK - Corpus Brown (Version XML TEI): http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml
- TXM (format pivot) : https://sourceforge.net/apps/mediawiki/txm/index.php?title=XML-TXM

  Les sources TEI sont pré-traitées par différentes feuilles de style XSL de la bibliothèque de TXM située à l'adresse : http://sourceforge.net/projects/txm/files/library/xsl

Langues des corpus traités

TXM traite tout corpus encodé en Unicode, y compris avec système d'écriture de droite à gauche.

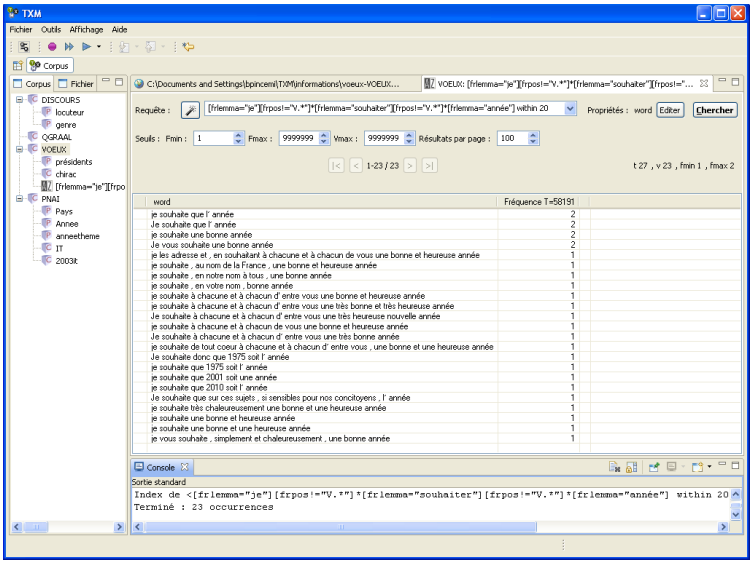
La prise en charge des langues au niveau des mots est assurée par les outils de TAL. Par exemple TreeTagger permet de travailler avec les lemmes des langues suivantes : BG, DE, EN, ES, ET, FR, FRO, FRP, GL, IT, LA, PT, RU, SW, ZH.

Documentation

- Page principale sur le site du projet Textométrie : http://textometrie.ens-lyon.fr/spip.php?article98&lang=en
- Site wiki de la communauté francophone des utilisateurs de TXM à l'adresse https://listes.cru.fr/wiki/txm-users (comprend une FAQ)
- Toute la documentation en ligne de TXM : http://sourceforge.net/projects/txm/files/documentation

Assistance utilisateurs

L'assistance est assurée par le biais de deux listes de discussion (voir ci-dessous) et par un système de tickets de retour de bugs et de demandes de fonctionnalités : https://forge.cbp.ens-lyon.fr/redmine/projects/txm/issues



Liste de fréquences d'un patron de mots

Communauté des utilisateurs de TXM

La communauté est animée par le biais de deux listes de diffusion et d'un site wiki :

- *Liste francophone* : txm-users AT groupes.renater.fr (la plus active)
  - Archives publiques situées à l'adresse https://listes.cru.fr/sympa/arc/txm-users
- *Liste internationale* : txm-open AT lists.sourceforge.net
  - Archives : http://sourceforge.net/mailarchive/forum.php?forum_name=txm-open
- Le wiki des utilisateurs (en français) à https://listes.cru.fr/wiki/txm-users

Des stages de formation à TXM sont dispensés gratuitement environ tous les mois dans les locaux du laboratoire (https://groupes.renater.fr/wiki/txm-users/public/ateliers_txm) et des ateliers de formation payants sont régulièrement organisés dans le cadre de conférences (comme par exemple Digital Humanities en 2013 au Nebraska : http://dh2013.unl.edu/schedule-and-events/workshops/#TXM).

Contrats

- Jan 2007 – Déc 2010: S. Heiden, projet ANR Textométrie – lancement de la plateforme, Agence Nationale de la Recherche (ANR) contrat #ANR-06-CORP-029;
- Jan 2012 – Déc 2014: D. Peschanski, Équipex Matrice – développement de TXM pour historiens, contrat ANR #ANR-10-EQPX-21-01.

Mai 2014