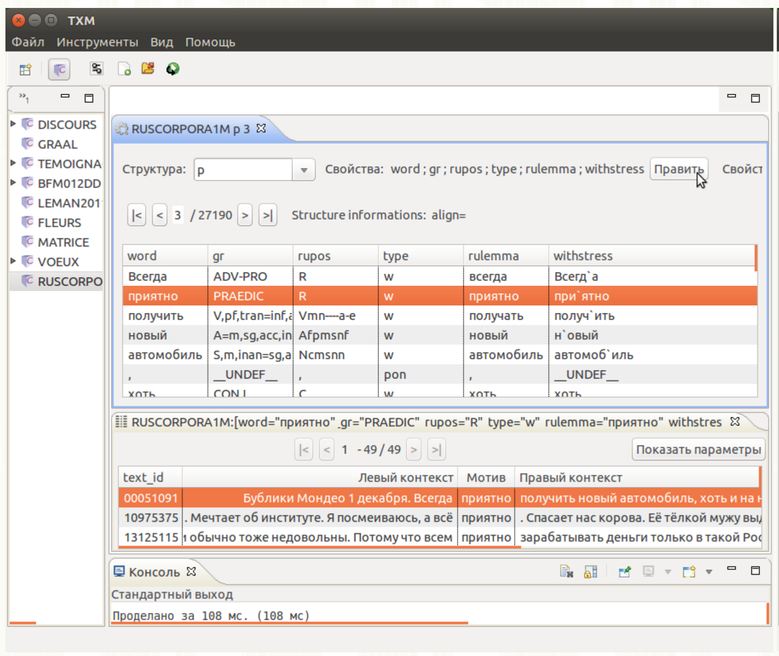
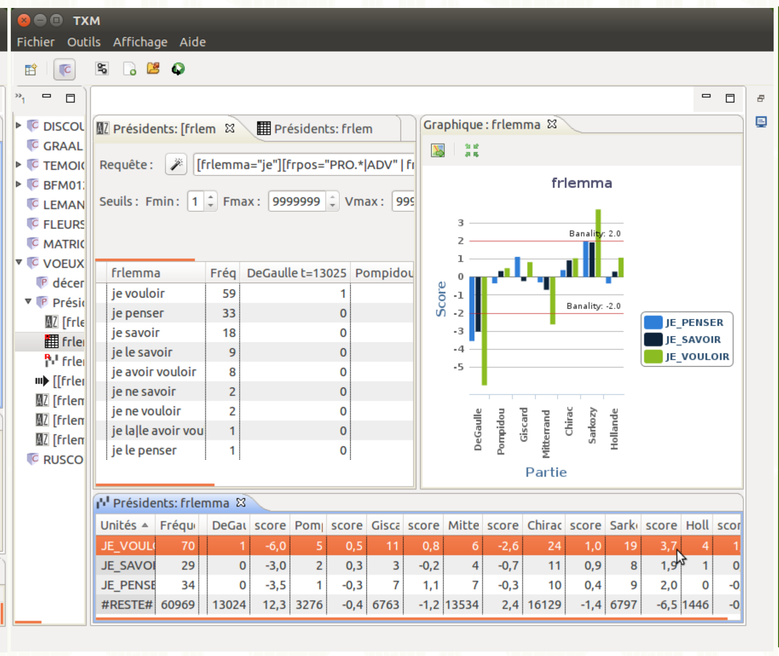
TXM, logiciel open-source d'analyse de corpus textuels

http://textometrie.ens-lyon.fr — Équipe de recherche CACTUS





Qu'est-ce que la textométrie ?

La textométrie (lexicométrie, statistique textuelle) propose une méthodologie et des techniques pour une analyse de corpus :

- ***qualitative** : observation en contexte dans les documents sources*
- ***quantitative** (répétitions, distributions) théoriquement fondée*
- ***endogène**, fondée d'abord sur la contextualisation dans le corpus et dans les unités textuelles*
- ***robuste**, pour toutes sortes de corpus : écrit comme oral transcrit, pour de nombreuses langues*
- ***semi-automatique** : le chercheur garde pleinement la conduite de l'analyse et de l'interprétation des résultats*
- ***exploratoire** : détection de régularités inaperçues dans une lecture et une analyse traditionnelle*

TXM dans une autre langue que le français (corpus et interface) : illustration de TXM en russe sur un échantillon de corpus gracieusement fourni par l'équipe du Corpus National Russe (http://ruscorpora.ru)

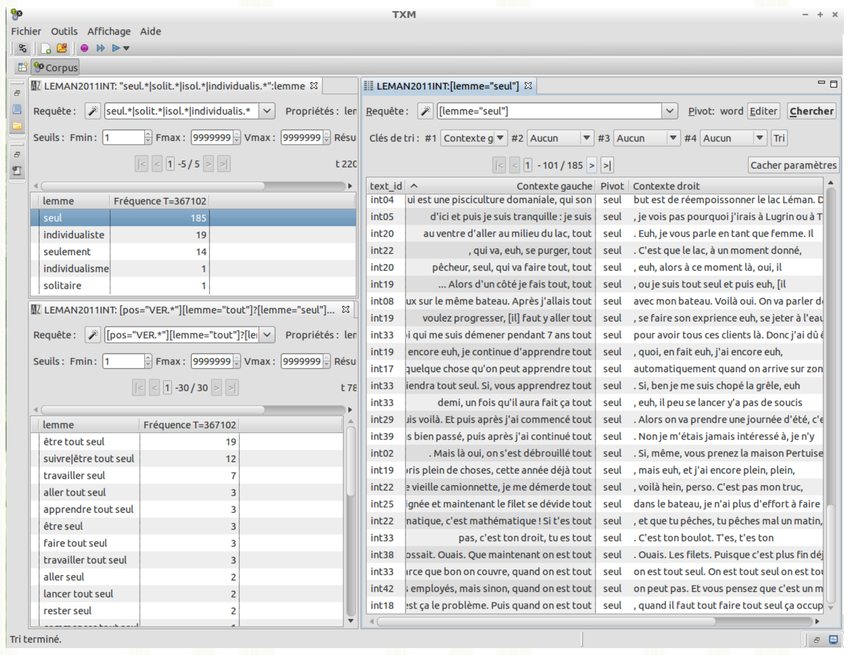
Exemple de calcul de spécificité sur des motifs : on observe que la construction « JE … VOULOIR » est caractéristique de N. Sarkozy dans le corpus VŒUX (J.-M. Leblanc, U. Paris 12).

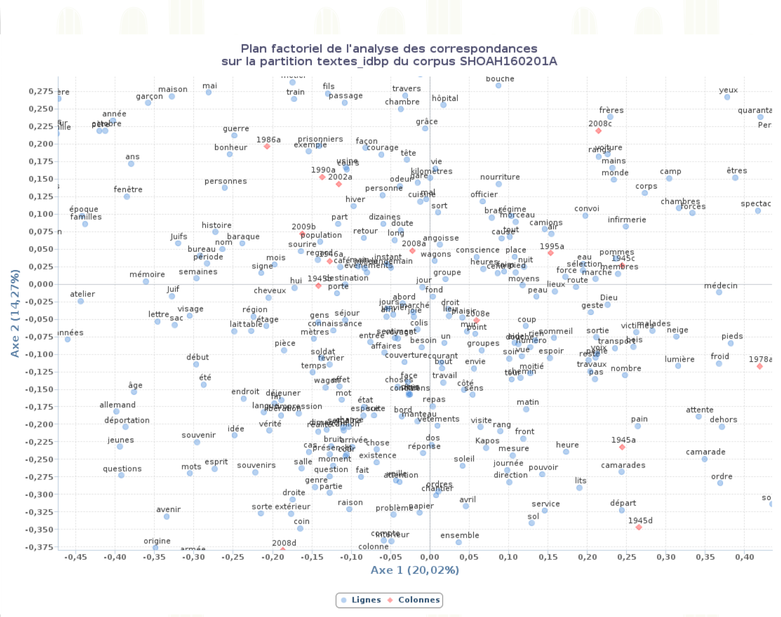
Recherche

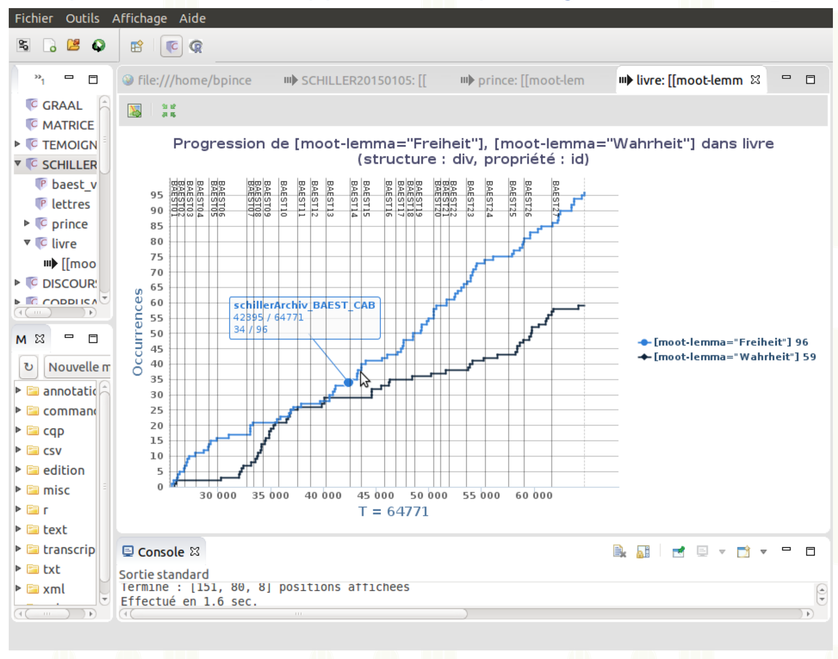
- **Modélisation** textométrique des textes et **philologie numérique**
- Typologie des **fonctionnalités** textométriques
- Modèles génériques de **mesures** pour la caractérisation quantitative des textes
- Éléments **méthodologiques** : points d'entrée et parcours d'analyse, herméneutique des sorties numériques, sémiotique des visualisations graphiques
- Linguistique et textométrie : liens avec la **sémantique interprétative**, épistémologie (travaux d'Étienne Brunet)

Innovation

- Traitement de trois paradigmes de corpus :
  - **Écrit** : du texte brut au texte structuré (standards **Unicode**, **XML**, **TEI**)
  - **Transcriptions** d'enregistrements : texte synchronisé
  - **Corpus parallèles** : textes alignés
- Souplesse de l'import (multiples formats et personnalisation possible par script)
- Articulation avec le **TAL** (lemmatisation optionnelle à la volée avec TreeTagger)
- Moteur de recherche plein texte **CQP** : pour le repérage et le décompte d'unités linguistiques complexes
- Extension et généralisation des calculs textométriques pour le traitement des corpus structurés et étiquetés
- Environnement statistique **R** dont développement du package `textometry`







INDEX et CONCORDANCE : formulations et contextes de la solitude du pêcheur dans le corpus LEMAN (Y. Le Lay & al., EVS, ENS de Lyon).

AFC sur le corpus SHOAH (FMS & Équipex Matrice). Illustre la possibilité de « tailler » finement le tableau de données (ici neutralisation des variations graphiques bloc/Block et SS/S.S.) et de zoomer dans le graphique.

PROGRESSION au fil des chapitres de « liberté » et « vérité » (corpus SCHILLER, A. Lagny, IHRIM, ENS de Lyon).

Rayonnement interdisciplinaire

Utilisation dans de multiples disciplines des **SHS**, par exemple sur le site de Lyon :

- **Littérature, Philosophie (UMR 5317 IHRIM)** : éditions numériques outillées des *Lettres esthétiques* de Schiller, des écrits de Tchitcherine, des dossiers de *Bouvard et Pécuchet* de Flaubert ;
- **Histoire (UMR 5190 LARHRA)** : outil d'annotation sémantique en lien avec la plateforme SyMoGIH et le projet Bibliothèque Historique de l'Education ;
- **Géographie (UMR 5600 EVS)** : analyse de données textuelles (presse, enquêtes, documents administratifs) et recherche sur texte & SIG ;
- **Linguistique, Didactique (UMR 5191 ICAR)** : interface d'interrogation de la Base de français médiéval, analyse de transcriptions de séances de classe, édition numérique outillée d'un journal de guerre.

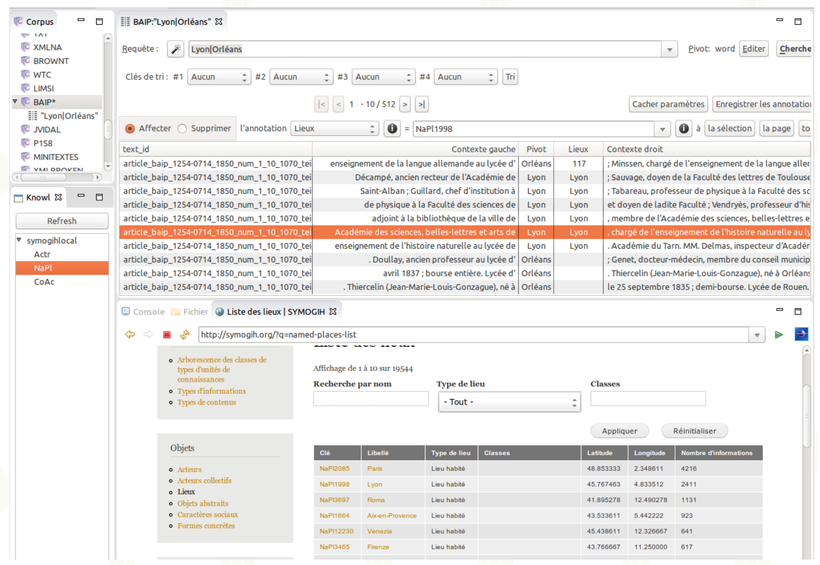
Diffusion du logiciel

- Open-source : plateforme modulaire, architecture standard, **développement mutualisé** (par ex. Besançon développe la nouvelle version du moteur graphique)
- Téléchargeable gratuitement : `http://sourceforge.net/projects/txm` 400 téléchargements par mois, la moitié à l'international
- Version pour poste (**Windows**, **Mac OS X**, **Linux**) et portail **Web** (ex. BFM)
- **Communautés** d'utilisateurs et de développeurs communiquant à travers des listes électroniques, des wikis, des sites web (site du projet, site de développement et de diffusion) ; de nombreuses **ressources en ligne** (corpus, documentation, tutoriels vidéo, etc.)

Perspectives

- **Corpus mutable et annotation dynamique** : possibilité de corriger et d'enrichir le corpus à travers les vues d'analyse, lien avec des référentiels type ontologie web sémantique et SIG.
- **Corpus multimodaux** : retour au document-source image (ex. manuscrit), audio (ex. enregistrement d'entretien), vidéo (ex. archive audiovisuelle)
- **Corpus diachroniques :** ajout et enrichissement de fonctionnalités d'analyse spécialisées
- Extension des **fonctionnalités**, **visualisations** graphiques interactives, etc.

Prototype de la fonctionnalité d'annotation, sur un corpus du projet BHE et avec le référentiel sémantique SyMoGIH (projets coord. par E. Picard, LARHRA-LLE, N. Fargier, Persée, F. Beretta, LARHRA, ENS de Lyon).



Repères temporels

- *Années 80-90 : **laboratoire de Saint-Cloud** « Lexicométrie et textes politiques », initiant le logiciel Lexico puis produisant le Lexploreur / Weblex*
- *2000 : le laboratoire de Saint-Cloud déménage avec l'**ENS à Lyon***
- *2007-2010 : **projet ANR Textométrie** « Fédération des recherches et développements en textométrie autour de la création d'une plateforme logicielle ouverte » (Lyon, Paris, Nice, Besançon)*
- ***2009 : début du développement** de TXM*
- *2012-2014 : développement et diffusion de TXM par l'**Équipex Matrice***

École Normale Supérieure de Lyon / 15, Parvis René Descartes F-69342 Lyon BP 7000 Cedex 07