3

CARACTERISTIQUES FONCTIONNELLES

Multiples formats de corpus : texte brut (,txt, copier/coller, Unicode) et structuré (XML, TEI), sorties de logiciels de préparation ou de sources de corpus (Cordial, Transcriber, Factiva), formats de logiciels d'analyse de corpus (Hyperbase, Alceste) et possibilité de modification ou d'ajout de scripts pour imports spécifiques

Ajout de métadonnées de texte par tableau (Excel ou autre) au format .csv

Étiquetage morphosyntaxique et une lemmatisation à la volée lors de l'import (TreeTagger, TnT)

Analyse textométrique de corpus « bruts » et de corpus structurés et multi-annotés, en exploitant souplement et en croisant les différentes annotations

Exports : tables au format .csv lisible par un tableur (Excel, Calc) et au format R ; figures et graphiques en format vectoriel .svg ; corpus compilé pour TXM ou au format XML TEI-TXM

Recherche de motifs linguistiques élaborés, combinant de multiples informations, sur plusieurs mots, avec des discontinuités et des variantes

Paramétrage très fin des concordances : composition des références de localisation, tris sur tout type d'information et découplés de l'affichage (exemple : tri sur le lemme et l'étiquette morphosyntaxique et affichage de la graphie)

Possibilité de construire, d'ajuster et de structurer le corpus de façon dynamique, au fil de l'analyse (les limites du corpus et les partitions ne sont pas obligatoirement uniques ni prédéfinies au moment de l'import)

Interface TXM RCP permettant l'affichage simultané et la mise en regard des résultats de différents calculs

Éditeur de texte intégré TXM RCP permettant l'édition de scripts, de corpus sources ou de paramètres d'importation

Gestion d'inscription et de comptes utilisateurs, paramétrage fin des contrôles d'accès aux textes TXM WEB

4

POUR L'ANALYSE TEXTOMETRIQUE

Les fonctionnalités sont implémentées dans des modules de la « boîte à outils » de la plateforme. Les interfaces des versions bureau **TXM RCP** et portail **TXM WEB** étant différentes, certaines fonctionnalités n'existent que sur l'une ou l'autre version et les paramétrages peuvent être plus ou moins nombreux. En voici une liste :

AFC : analyse factorielle des correspondances, pour une cartographie du corpus

CLASSIFICATION TXM RCP : plan factoriel

CONCORDANCE TXM RCP TXM WEB : affichage des contextes d'un mot ou motif donné, sous forme de concordances KWIC triables

CONTEXTES TXM WEB : affichage des contextes d'un mot ou motif donné sous forme de relevé d'extraits

COOCCURRENCES : mots statistiquement associés à un mot ou motif donné et apparaissant dans son voisinage

DESCRIPTION TXM RCP : récapitulatif des annotations et structures disponibles pour l'analyse

DIMENSIONS TXM WEB : dimension du corpus

ÉDITION : consultation des textes et outils de navigation

INDEX : liste des différentes réalisations d'un motif avec leur fréquence (ex. : tous les adjectifs qui suivent le mot « métier ») ; construction de tableaux croisant d'une part des mots/motifs, et d'autre part des parties d'un corpus, avec leur fréquences locales (TABLE LEXICALE TXM RCP)

LEXIQUE : liste des valeurs attestées pour une propriété, avec leur fréquence, par ex. liste des mots (formes graphiques ou lemmes), liste des étiquettes morphosyntaxiques

PARTITIONS : division du corpus en parties à contraster entre elles (voir aussi SOUS-CORPUS)

PROGRESSION TXM RCP : graphique visualisant la position des occurrences d'un mot/motif au fil du déroulement d'un corpus

RÉFÉRENCES : construction automatique d'index listant les localisations d'apparition d'un mot/motif

SOUS-CORPUS : la construction de sous-corpus par sélection de textes TXM WEB offre une interface très riche pour construire des corpus de travail à partir d'une collection de textes caractérisés par de nombreuses métadonnées (auteur, titre, date d'écriture, date de publication, genre textuel...)

SPÉCIFICITÉS : mots ou étiquettes (ex. catégories grammaticales) caractéristiques d'une partie ou d'un sous-corpus par rapport à l'ensemble du corpus, profil de répartition d'un ou plusieurs mots/motifs sur différentes parties et affichage sous forme de diagramme en bâtons

5

(ajouter exemples/références de réalisations avec TXM)

2

QU'EST-CE QUE LA TEXTOMETRIE ?

Elle se développe en France à partir des années 1970, dans la lignée des recherches pionnières en statistique lexicale. Elle reprend et poursuit également les méthodes d'analyse des données (analyses factorielles, classifications) appliquées aux données linguistiques et permet ainsi de générer des cartographies synthétiques et visuelles des mots et des textes tels qu'ils s'apparentent ou s'opposent au sein d'un corpus. Les résultats des calculs sont des réorganisations synthétiques, sélectives et suggestives : listes ordonnées, visualisation cartographiques, regroupements, mises en valeur au fil du texte. L'interprétation des calculs se fonde sur des indicateurs chiffrés mais aussi sur l'examen systématique des contextes. La textométrie est appréciée dans des disciplines très diverses : archives historiques, dépouillement d'enquêtes avec questions ouvertes, œuvres littéraires, etc. Elle permet une observation à la fois fine et globale des textes, et donc une exploitation relativement complète des données rassemblées dans ces corpus. La plateforme TXM a été conçue pour reprendre la tradition lexicométrique (implémentée par Hyperbase, Lexico 3...) avec des corpus enrichis et structurés. Elle a été initiée dans le cadre du projet ANR Textométrie (2007-2010) et continue son développement grâce à son réseau de partenaires et au soutien de l'Equipex Matrice (2012-2014).

6

LIENS UTILES

le projet Textométrie
http://textometrie.ens-lyon.fr **ou QR code**

pour télécharger et développer TXM
http://sourceforge.net/projects/txm/files/software

la documentation en PDF
http://sourceforge.net/projects/txm/files/documentation/Manuel%20de%20Reference%20TXM%200.5_FR.pdf/download
(à remplacer par lien vers v0.7?)

des exemples de corpus
https://sourceforge.net/projects/txm/files/corpora

démo de la version en ligne TXM WEB ??

la liste de discussion des utilisateurs txm-users
https://groupes.renater.fr/sympa/info/txm-users

le wiki des utilisateurs de TXM
https://groupes.renater.fr/wiki/txm-users (fr)

le wiki des développeurs de TXM
http://sourceforge.net/apps/mediawiki/txm (en)

la fiche PLUME de TXM
https://www.projet-plume.org/relier/txm

la page TXM du wiki de la TEI
http://wiki.tei-c.org/inde.php/TXM

(+logo) LA PLATEFORME LOGICIELLE TXM

PRESENTATION ET INITIATION

TXM, c'est quoi ?

✔ un logiciel libre : sous licence open-source GPL v3, téléchargeable gratuitement en ligne, offrant la possibilité de participer à son développement

✔ un outil multiplateforme :

Version TXM RCP	Windows	Mac OS X	Linux	Moteur de recherche CWB CQP
Version TXM WEB	À installer sur un serveur et interrogeable en ligne avec navigateur internet (technologie GWT)			Moteurs de recherche CWB CQP et TigerSearch

✔ intègre l'environnement R pour les bibliothèques de calculs statistiques

✔ langages de développement : Java et C ; scripts Groovy et R

POUR NOUS CONTACTER ***(+ num. tél et e-mail)***

Equipe projet Textométrie :

Bénédicte PINCEMIN, conception fonctionnalités, méthodologie
Serge HEIDEN, responsable du projet
Matthieu DECORDE, développeur principal

Laboratoire ICAR – UMR 5191 CNRS
ENS site Descartes, Lyon 7e Métro B Debourg

textometrie@ens-lyon.fr