Atelier TEI ++ : TP 8b : Plateforme TXM

Alexei Lavrentiev, 22/11/2012

1 Objectifs de l'exercice

Dans cette présentation nous allons :

- apprendre à quoi sert la plateforme TXM
- explorer les fonctionnalités du logiciel sur des corpus exemples
- tester la procédure d'importation de corpus dans TXM

2 Plan

1. Présentation de la plateforme TXM
2. (Installation du logiciel)
3. Analyse du corpus GRAAL (fourni avec TXM)
4. Chargement et analyse du corpus BROWN "pré-compilé"
5. Importation et analyse d'un corpus de test (extraits de Molière, du Grand Cyrus et de l'Astrée)

Vous aurez besoin du logiciel TXM 0.7 beta 3 et des fichiers fournis dans votre dossier TXM. Si vous n'avez pas installé TXM avant cet atelier vous pourrez interroger les même corpus sur le portail en ligne.

3 Plateforme TXM

- 2007-2010 Impulsion projet ANR
  - 6 laboratoires partenaires (4 FR, 1 UK, 1 CA)
  - TXM Windows, Mac & Linux application de bureau
  - Portail TXM : accès en ligne (inscription, contrôle d'accès)
- Performance : Milliards de mots (théorique)
  - Composants open-source clés : R (stat.) & CQP (moteur rech.)
  - Outils de TAL : taggers
  - Unicode (TXT)
  - XML
  - compatible XML-TEI

4 Plateforme TXM : liens et contacts

- Pour télécharger la dernière version : http://sf.net/projects/txm
- Pour explorer des corpus en ligne : http://txm.risc.cnrs.fr/demo
- Pour s'inscrire à une formation, poser une question à l'équipe : textometrie@ens-lyon.fr
- Pour être informé de l'évolution du logiciel et échanger avec la communauté des utilisateurs, inscrivez-vous à la liste txm-users@groupes.renater.fr (https://groupes.renater.fr/sympa/info/txm-users)
- Pour en savoir plus sur le projet : http://textometrie.ens-lyon.fr

5 Outils pour l'Analyse de Contenu dans TXM

Analyse qualitative

- Concordances Kwic
- Lecture & Navigation dans les éditions

Analyse quantitative

- Listes de fréquences
- Cooccurrents statistiques
- Mots spécifiques à un sous-corpus
- Analyses factorielles, Classification
- Construction de Sous-corpus et Partitions

6 Catégories de données dans TXM

- Unités lexicales (mots simples ou composés)
  - Propriétés : lemme, partie du discours...
- Unités textuelles (livre, article...)
  - Propriétés = métadonnées principales (auteur, titre, date, domaine, genre…)
- Structure interne (phrase, paragraphe, section...)
  - Propriétés = métadonnées secondaires
- Plans textuels
  - Corps du texte/Titres, Discours direct, Prises de parole
  - Langue principale (par ex. français…), lange secondaire (par ex. latin…)
- Édition
  - Pagination (sauts de page)
  - Mise en forme (styles)
- Hors-texte : commentaires, <teiHeader>, apparat critique
- Alignement (corpus parallèles)

7 Déroulement de la séance pratique

1. Démonstration des fonctionnalités principales sur les trois corpus choisis
2. Travail pratique individuel sur l'un des corpus au choix:
   - Graal (édition d'un texte avec un riche balisage TEI, ancien français)
   - Brown (corpus équilibré, approprié aux analyses statistiques, anglais)
   - CEPM (procédure d'importation, petits échantillons d'éditions TEI différentes, français classique)

8 Exploration du corpus GRAAL 1

- Lancez TXM
- Si vous voulez changer la langue de l'interface, utilisez le menu File / Change Language
- Faites un clic-droit sur le nom du corpus GRAAL
  - Sur un Mac, vous pouvez : faire un "Pomme + clic" ou faire un simple clic et utiliser les icones situées sous le menu principal en haut.
  - Vous verrez les principales caractéristiques du corpus : le nombre de mots, le nombre et une sélection de valeurs des propriétés lexicales et de structures.

- Une grande partie de ces informations provient du balisage TEI du texte source !

Vous pouvez aussi interroger le corpus Graal sur le portail DEMO

9 Exploration du corpus GRAAL 2

Recherche "panoramique"

- Faites un clic-droit sur le nom du corpus GRAAL
- Dans le menu contextuel, sélectionnez la fonction "Lexique", puis la propriété "pos" (étiquette morphosyntaxique Cattex 2009) et cliquez sur "OK".
  - Vous verrez la liste des étiquettes morphosyntaxiques de ce texte classée par ordre de fréquence.
- Double-cliquez sur une ligne du lexique (par exemple "INJ" = interjection)
  - Une concordance des formes correspondantes à l'étiquette choisie s'ouvrira dans un nouvel onglet
- Double-cliquez sur une ligne de la concordance
  - Une vue "Édition" s'ouvrira au-dessus de la concordance.
  - L'occurrence correspondante à la ligne de la concordance sélectionnée sera surlignée.

10 Exploration du corpus GRAAL 3

Recherche ciblée

- Cliquez à nouveau sur le nom du corpus GRAAL
- Dans le menu contextuel, sélectionnez la fonction "Index"
- Dans le formulaire qui s'affiche, tapez une forme que vous souhaitez rechercher, par exemple "Lancelot"
- Vous obtenez le nombre d'occurrences de la forme recherchée
- Comme pour le lexique, vous pouvez accédez aux concordances et à l'édition du texte en faisant des double clics.

Le moteur de recherche CQP utilisé par TXM permet de formuler des requêtes très sophistiquées portant sur les propriétés de mots individuels (formes et annotations), sur des successions de mots ou sur des structures (éléments XML de la source)

11 Exploration du corpus GRAAL 4

Progression

- Revenez à l'onglet Index
- Pour obtenir l'index des noms propres, tapez la requête suivante :
  - [pos="NOMpro"]
- Sélectionnez les 3 premières lignes (Ctrl + Clic ou Maj + ↓)
- Faites un clic-droit et sélectionnez "Envoyer à la Progression" dans le menu contextuel
- Cliquez sur OK dans le formulaire qui s'affiche
- La ligne du graphique "monte" à chaque occurrence du nom du personnage correspondant

12 Exploration du corpus BROWN

- Dans TXM, sélectionnez l'action "Fichier / Charger", ouvrez le fichier brown-bin.zip dans le dossier Travaux/txm
- L'opération peut prendre quelques minutes
- Faites un clic-droit sur le nom du corpus BROWN et sélectionnez la fonction "Créer une Partition"
- Tapez un nom de la partition, par exemple "genres"
- Assurez-vous que la structure sélectionnée est "text" et sélectionnez la propriété "type"
- La nouvelle partition doit apparaître sous le nom du corpus BROWN
- Faites un clic-droit sur le nom de la partition et cliquez sur "Dimensions". Vous verrez un histogramme de la taille des parties du corpus en nombre de mots
- Sur la même partition, sélectionnez la fonction "AFC", puis la propriété "enpos" et cliquez sur OK
- Vous verrez le plan factoriel des genres du corpus selon la fréquence

Pour en savoir plus sur le corpus Brown : http://icame.uib.no/brown/bcm.html

La version TEI du corpus Brown a été produite par Lou Burnard et est accessible dans la bibliothèque NLTK http://nltk.org/nltk_data

Le corpus BROWN peut être interrogé sur le portail http://txm.bfm-corpus.org/demo

13 Importation du corpus CEPM 1

Le corpus est composé de petits échantillons de textes des CEPM http://cepm.paris-sorbonne.fr fournis par A. Geffen :

- *L'Astrée* (1607-1627) d'Honoré d'Urfé
- *Artamène ou le Grand Cyrus* (1649-1653) de Madeleine de Scudéry, http://www.artamene.org
- *Le Malade imaginaire* (édition 1674) de Molière

Avant d'importer les fichiers, nous avons effectué queleques modifications :

- conversion à la TEI P5 (à l'aide de la feuille de style fournie avec oXygen)
- quelques petites corrections pour valider les documents (facultatif)
- extraction des métadonnées (feuille de style metadata-cepm.xsl)
- filtrage d'adaptation au format compatible TXM (feuille de style filter-tei2xmlwtxm.xsl, appliquée lors de l'import)

14 Importation du corpus CEPM 2

Plusieurs modules d'importation de corpus sont proposés dans TXM...

Les modules "TEI" ne prennent pour le moment en charge que les schémas personnalisés de la BFM, du Frantext et de la BFM.

En attendant la mise en place d'un module TEI plus générique, nous pouvons utiliser le module "xml/w+csv", qui permet d'importer n'importe quel document xml. Dans ce module :

- Chaque fichier constitue une unité textuelle

- Les métadonnées des unités textuelles sont fournies dans un tableau CSV où la première colonne "id" indique le nom du fichier sans extension
- Tous les éléments XML deviennent des structures indexées dans TXM
- Tous les nœuds text() sont tokenisés et inclus dans le corpus (y compris <teiHeader>, <note>, <choice>...)
- Vous pouvez utiliser une feuille xsl pour "filtrer" les éléments et les parties du texte que vous ne voulez pas indexer
  - Attention : pour le moment, il faut que tous les mots soient annotés avec les mêmes attributs, sinon il y a un risque de confusion des index. Ce bug sera résolu dans la verision 0.7 stable.

15 Importation du corpus CEPM 3

Dans TXM

- Utilisez le menu "Fichier / Importer / XML/w+CSV"
- Dans le formulaire des paramètres d'import, sélectionnez le dossier des sources "Travaux/txm/cepm"
- Cliquez sur "Langue principale" et sélectionnez l'option "Deviner"
- Cliquez sur "Feuille XSL d'entrée" et sélectionnez le fichier "Travaux/txm/filter-tei2xmlwtxm.xsl"
- Cliquez sur le bouton vert en haut du formulaire
- Le processus d'importation devrait prendre 2-3 minutes. Quand il sera terminé, un nouveau corpus "CEPM" apparaitra dans la barre de navigation à gauche.

Vous pouvez maintenant tester les fonctionnalités TXM sur ce micro-corpus ou encore travailler sur une version plus complète (contenant le texte intégral du Grand Cyrus)

- en téléchargeant le corpus "binaire" : http://partage-fichiers.ens-lyon.fr/89xpugl2z1 (128 Mo) ou
- en vous connectant au portail http://txm.bfm-corpus.org/demo avec l'identifiant "CepmTester" et le mot de passe "tei20121122"