Tutoriel de préparation et d'import de corpus dans TXM
-
Philologie numérique progressive avec TXM

Serge Heiden et Alexei Lavrentev
2014-2017

Ce document présente une séquence type de l'atelier de formation « préparation de corpus et import dans TXM » ainsi que les sources du corpus VOEUX sur lesquelles il s'appuie.

Table des matières

N° d'édition : 96
Contenu : 10 pp., 3286 occ., 0 ill., 0 tab.
Date d'édition : 22/03/17, 19:44:27

Introduction à l'environnement d'importation de TXM

Cette section n'est pas rédigée, voir le document « Diapositives - Atelier preparation de corpus et import dans TXM.pdf » dont voici le sommaire :

- Présentation des trois grandes catégories de corpus analysables avec TXM :
  - A. Corpus de **textes écrits**, comprenant éventuellement des éditions paginées incluant les images de fac-similés (comme par exemple de manuscrits médiévaux, d'auteur ou encore d'enfants)
  - B. Corpus de **transcriptions d'enregistrements**, éventuellement synchronisées avec l'audio ou la vidéo
  - C. Corpus **multilingues alignés** au niveau d'une structure textuelle comme la phrase ou le paragraphe
- présentation des 4 niveaux principaux de représentations textuelles importés par TXM (propriétaire, TXT, XML, TEI) et du schéma de normalisation associé
- présentation des éléments composant un corpus pour TXM - le modèle de données de TXM :
  - textes (unités textuelles) / leurs métadonnées
    - sections (structures textuelles) / leurs propriétés
      - mots (unités lexicales) / leurs propriétés
  - hors texte et plans textuels
  - éditions de texte
  - etc.
- relation entre ce modèle et les différents niveaux d'import : carte croisant les niveaux de représentation avec les éléments du modèle de données disponibles dans TXM.

Import 0 - import à partir du presse-papier

L'import presse-papier est l'import nécessitant le moins de préparation et le plus rapide pour analyser un « texte » dans TXM.

- dans n'importe quel logiciel :
  - copier un contenu textuel (d'un document de traitement de texte, d'une page web ou d'un mail, etc.)
- dans TXM :
  - commande 'Fichier / Importer / Presse-papier' pour lancer l'import
  - puis Lexique, double-clic pour concordance, double-clic pour retour au texte avec mise en évidence du mot
  - vérifier le contenu de l'édition
  - présenter les propriétés de mots
  - réglages de l'étiquetage par TreeTagger dans les préférences

Retour sur la carte des niveaux de représentation.

Utilisation d'un extrait des sources du corpus VOEUX

La suite de l'atelier s'appuie sur un corpus limité à trois textes :

- t0015, Pompidou, 1973
- t0022, Giscard, 1980
- t0036, Mitterrand, 1994

Ce corpus est un extrait des sources du corpus VOEUX utilisé lors de l'atelier d'initiation à TXM, généralement suivi avant l'atelier de préparation de corpus et d'import dans TXM, et gracieusement fourni par Jean-Marc Leblanc de l'EA 3119 CÉDITEC de l'Université Paris Est. Les sources sont fournies sous différentes représentations (selon différents formats) pour un apprentissage progressif de l'importation de sources dans TXM.

Les module d'import XTZ sera utilisé avec les corpus *Tour du monde en 80 jours* de Jules Verne et un extrait *Des cas des nobles hommes et femmes* de Boccace traduit en français par Laurent de Premierfait au XV^{e} siècle. Le premier corpus a été préparé par Serge Heiden à partir de Wikisource. Le second a été préparé par Alexei Lavrentev à partir d'une transcription du manuscrit BnF fr. 226 réalisée par Céline Barbance-Guillot.

Philologie numérique progressive avec TXM

La taille limitée du corpus extrait permet d'importer rapidement de multiples fois les mêmes sources dans TXM à différents niveaux d'encodage produisant autant de versions différentes du même corpus à analyser. Ces niveaux sont conçus de façon progressive, du texte brut le plus simple à un encodage XML plus complet, de sorte à ce que chaque nouvel élément d'encodage des sources offre une possibilité d'exploitation supplémentaire dans TXM. Il s'agit de présenter les capacités adaptatives de TXM pour l'exploitation de sources de différents niveaux de représentation et d'encodage. Ce qui offre à l'utilisateur la possibilité de choisir au mieux le niveau, et donc l'effort, de préparation et d'encodage de ses sources en fonction de ses besoins d'exploitation.

Progressif ne veut pas dire linéaire : la séquence présentée dans cet atelier est modulable, la progression peut passer directement à certaines étapes en fonction de l'effort et des informations d'encodage disponibles et de ce qui est le plus utile à l'exploitation du corpus. Une première exploitation du corpus à un niveau de représentation et d'encodage donnés, peut justifier un nouvel effort d'encodage du corpus pour une exploitation plus fine ou avancée, formant un cycle philologique complet d'encodage-exploitation ou établissement de texte-interprétation.

Une session d'atelier comporte typiquement une progression d'importations pour les trois éléments fondamentaux du modèle de corpus de TXM :

- les textes (ou unités textuelles) et leurs métadonnées
  - les sections (ou structures textuelles) et leurs propriétés
    - les mots (ou unités lexicales) et leurs propriétés

L'encodage TEI[1] et les encodages spécialisés (transcriptions d'enregistrements[2], corpus alignés[3]) font l'objet de séances complémentaires.

De nombreux paramètres d'importation supplémentaires sont disponibles. Ils sont documentés dans le manuel de TXM.

Données exemples utilisées par l'atelier

Une archive de sources sous licence libre est fournie comme support de l'atelier.

Les trois textes du corpus VOEUX utilisés sont fournis sous trois formats :

- le répertoire 'voeux-odt' contient les textes au format LibreOffice *Writer* Open Document

1 Voir les « Tutoriels d'importation de corpus de textes XML-TEI » : https://groupes.renater.fr/wiki/txm-users/public/tutoriels_import_xml-tei.
2 Voir le « tutoriel ».
3 Voir la documentation du module d'import XML-TMX.

Text (.odt) - équivalent de Microsoft *Word* (.doc)

- le répertoire 'voeux-txt' contient les textes au format texte brut (.txt) en Unicode UTF-8
- le répertoire 'voeux-xml' contient les textes au format XML (.xml)
- le répertoire 'voeux-xml-full' contient les textes au format XML (.xml) entièrement encodés

Les métadonnées des textes sont fournies en trois formats :

- 'metadata.csv' au format CSV (.csv) – virgule « , » comme caractère séparateur de colonnes et champs de texte délimités par une double apostrophe droite « " »
- 'metadata.ods' au format Open Document Spreadsheet (.ods) du logiciel LibreOffice *Calc*
- 'metadata.xls' au format '.xls' du logiciel Microsoft Office *Excel*

Il faut utiliser le fichier 'metadata.csv' avec TXM.

Déroulé type d'une séance d'atelier

La suite de ce document présente la succession type d'opérations de préparation de sources et d'import dans TXM en introduisant progressivement aux notions et en encodant progressivement les corpus avec de plus en plus d'informations.

Il n'est pas nécessaire de suivre toutes les étapes, notamment en fonction des capacités des apprenants et du temps disponible. On peut distinguer les aspects techniques suivants :

- notion de texte brut
- notion d'encodage de caractères
- notion d'encodage de sauts de ligne
- notion de texte XML
- usage de macros TXM
- éditeur de texte intégré
- navigateur de fichiers
- formulaire de paramètres d'import
- manipulations de formats de textes
- notion d'édition de texte par chercher/remplacer d'expressions régulières

Les deux derniers aspects peuvent être abordés de façon plus ou moins développée.

Pour illustrer les exploitations possibles des informations importées, l'atelier utilise les commandes suivantes : Description, Lexique, Index, Concordance, retour au texte, Sous-corpus, Partition.

Transformation 1 - Traitement de texte vers texte brut – TXT

Beaucoup de sources sont éditées avec des traitements de texte, LibreOffice *Writer* ou Microsoft *Word*. C'est également le format de sortie le plus fréquent des logiciels d'OCR.

Il existe diverses façons expérimentales de transformer en XML-TEI le contenu de tels documents. Dans ce cas l'import dans TXM se fait par des modules d'import de type XML-TEI qui sont présentés dans des ateliers complémentaires.

On peut également importer ces sources sous forme de texte brut. Pour cela il faut d'abord convertir les documents de traitement de texte en fichiers de texte brut.

A. Utiliser la commande « Enregistrer sous » des traitements de texte pour sauver les documents en texte brut, en indiquant d'utiliser l'encodage des caractères « Unicode[4] UTF-8 ».

B. Quand on a plusieurs documents à convertir d'un coup, la macro Text2TXT[5] de transformation par lot de fichiers au format traitement de texte vers du texte brut peut être utile.

Utilisation de la macro Text2TXT :

- inputDirectory : voeux-odt
- extension : odt

Import 1 - import d'un corpus de textes brut – TXT

Le premier import à réaliser après l'import Presse-papier consiste à importer un corpus de plusieurs textes brut (dans le cas d'un import presse-papier, le corpus ne peut contenir qu'un seul « texte ») :

- importer les textes du répertoire 'corpus/voeux-txt' avec le modèle de langue FR (garder le formulaire de paramètres ouvert) : commande 'Fichier / Importer / TXT+CSV'
  - désigner le répertoire source
  - choisir la langue du corpus
- analyser la Description du corpus et notamment les propriétés de mots fournies par TreeTagger
- visualiser les Lexique de word, frlemma et frpos
- parcourir le jeu d'étiquettes TreeTagger FR
- utiliser l'identité des textes pour calculer un sous-corpus ou une partition

Import 1.1 - Texte brut et métadonnées

L'import suivant consiste à associer des propriétés à chaque texte du corpus (appelées « métadonnées ») :

- copier le fichier 'voeux-metadata/metadata.csv' dans le répertoire 'voeux-txt'
- ré-importer le répertoire 'voeux-txt' (en utilisant le même formulaire d'import)
- analyser les métadonnées de texte dans la Description
- utiliser des métadonnées de textes pour calculer un sous-corpus ou une partition (et comprendre la différence dans les regroupements de textes)

Import 1.2 - Texte brut et nouvelles métadonnées

L'import suivant consiste à ajouter une métadonnée originale à chaque texte pour pouvoir l'exploiter dans TXM :

4 Introduire à Unicode si nécessaire.

5 La macro Text2TXT nécessite d'avoir LibreOffice ou OpenOffice installé sur la machine de l'utilisateur (jusqu'à TXM 0.7.5 inclus, il faut utiliser LibreOffice version 3.x. À partir de TXM 0.7.6, n'importe quelle version de LibreOffice convient).

- ajouter une nouvelle métadonnée à 'metadata.csv' (par exemple l'orientation politique 'pol' aux valeurs 'G' ou 'D')
- ré-importer le répertoire 'voeux-txt' (en utilisant le même formulaire d'import)
- vérifier les nouvelles métadonnées dans la Description
- utiliser la nouvelle métadonnée pour calculer un sous-corpus ou une partition

Transformation 2 - Texte brut vers XML

Pour pouvoir encoder des structures internes aux textes et des mots particuliers, il est nécessaire d'utiliser le format XML.

Il est facile de transformer des fichiers au format texte brut en XML (qui permet une exploitation plus riche des sources).

Il suffit :

- d'y remplacer tous les caractères '&' par '&amp ;'
- remplacer tous les caractères '<' par '<'
- ajouter une balise XML de début de fichier : par exemple '<discours>'
- ajouter une balise XML correspondante de fin de fichier : par exemple '</discours>'
- changer l'extension de fichier '.txt' en '.xml'

TXM permet d'éditer les sources avec son éditeur de texte intégré[6], voir sa documentation : http://txm.sourceforge.net/doc/manual/manual23.xhtml#toc64

On accède à l'éditeur de texte par « Fichier / Ouvrir... » (ou son icone) ou bien par l'explorateur de fichiers (« Affichage / Vues / Fichier »), navigation puis double-clic sur le fichier à éditer.

A. Utiliser l'éditeur de texte intégré pour transformer les fichiers texte brut en fichiers XML par rechercher/remplacer.

B. Quand on a plusieurs fichiers à convertir d'un coup, la macro TXT2XML de transformation de fichiers au format texte brut en format XML élémentaire peut être utile.

Utilisation de la macro TXT2XML :

- inputDirectory : voeux-txt
- encoding : UTF-8
- rootTag: discours

Import 2 - import d'un corpus de textes XML

Pour faire notre premier import XML, soit nous utilisons les fichiers XML que nous venons de construire, soit nous utilisons les fichiers XML déjà préparés du répertoire 'corpus/voeux-xml'.

- importer les textes du répertoire 'corpus/voeux-xml' avec le modèle de langue FR (on peut copier le fichier de métadonnées 'metadata.csv' dans ce répertoire de sources mais il ne sera plus utilisé dans les exercices)
- constater que le corpus est similaire au corpus de textes bruts modulo :

6 Une liste de logiciels d'édition de texte externes populaires est fournie en annexe.

- un formatage des pages d'édition légèrement différent
- la disparition de l'encodage automatique des phrases

Encodage 1 – Ajout d'une structure XML, le paragraphe

La balise XML ... permet d'encoder des paragraphes dans les textes.

On peut encoder facilement des paragraphes par chercher/remplacer d'expressions régulières à l'aide de l'éditeur de texte intégré de TXM. Les expressions régulières correspondent aux patrons de recherche des requêtes CQL.

A. Utiliser l'éditeur de texte intégré pour modifier les trois fichiers XML : 't0015.xml', 't0022.xml', 't0036.xml'

- chercher tous les '\.$\n'[7]
- et les remplacer par '. \n '
- finaliser l'encodage manuellement

B. Quand on a plusieurs fichiers sur lesquels appliquer un chercher/remplacer d'un coup, la macro SearchReplaceInDirectory peut être utile.

Remarque : les fichiers du répertoire 'corpus/voeux-xml' contiennent déjà des paragraphes.

Import 2.1 - Texte XML structuré

Il nous faut ré-importer le corpus pour pouvoir manipuler les paragraphes dans TXM :

- ré-importer les textes du répertoire 'corpus/voeux-xml'
- analyser la Description du corpus et notamment les structures
- calculer la concordance des premiers mots de paragraphes, aller à l'édition et vérifier la position des mots mis en évidence et le formatage des paragraphes

Encodage 1.1 – encodage d'une propriété de structure

Comme les métadonnées des textes, on peut associer des propriétés aux structures internes. Encoder les premiers paragraphes de chaque texte avec l'attribut « type='ouverture' » et les derniers paragraphes avec l'attribut « type='cloture' ».

Import 2.2 - Texte XML structuré annoté

Il nous faut ré-importer le corpus pour pouvoir manipuler les paragraphes d'un certain type dans TXM :

- ré-importer les textes du répertoire 'corpus/voeux-xml'
- analyser la Description du corpus et notamment les propriétés de la structure 'p'
- calculer la concordance des premiers mots de paragraphes de type 'ouverture', aller à l'édition et vérifier la position des mots mis en évidence
- calculer le sous-corpus du premier texte 't0015', puis le sous-corpus de son premier paragraphe, puis son lexique

7 Introduire les trois types de saut de ligne suivant les systèmes d'exploitation si nécessaire.

Import 2.3 - Texte XML en filtrant une structure

En ré-important en utilisant un filtre XML, on peut choisir d'éliminer certains types de paragraphes du corpus dès l'import :

- ré-importer les textes du répertoire 'corpus/voeux-xml' en ajoutant au formulaire de paramètres le filtre XSL ' filter-out-p.xsl' modifié pour filtrer les paragraphes d'ouverture
- vérifier que les premiers paragraphes ont bien été ignorés dans l'édition

Encodage 2 – pré-codage du mot « parce que »

Il est possible d'encoder la délimitation de certains mots avec une balise XML.

Identifier le texte contenant des « parce que » à l'aide d'une concordance de « parce.* ».

Éditer le fichier 't0022.xml' pour encoder tous les « parce que » et le « parce qu' » à l'aide d'une balise XML <w>...</w> :

- <w>parce que</w>
- <w>parce qu'</w>

Import 3 - Texte XML avec certains mots pré-codés

Il nous faut ré-importer pour bénéficier de ces nouvelles délimitations dans TXM :

- ré-importer les textes du répertoire 'corpus/voeux-xml'
- vérifier l'encodage des « parce que » : faire la concordance de « parce.* », et vérifier les retours au texte
- faire la concordance des mots contenant un espace : [word=".* .*"]

Encodage 2.1 – pré-codage d'une propriété de mot

Comme pour les métadonnées de textes et les propriétés de structures, il est possible d'associer des propriétés aux mots balisés.
Différencier le « parce qu' » (élidé) avec l'attribut « elision='yes' » :

- <w elision='yes'>parce qu'</w>

Import 3.1 - Texte XML avec certains mots pré-codés annotés

Il nous faut ré-importer pour bénéficier de ces nouveaux types de mots dans TXM :

- ré-importer les textes du répertoire 'corpus/voeux-xml'
- Vérifier la présence de la nouvelle propriété de mot dans la description du corpus
- faire la concordance des mots élidés () : [elision="yes"]
- constater qu'il n'y a pas de différence de traitement avec les propriétés « frlemma » ou « frpos » en faisant le lexique de la propriété 'elision' (tout en sachant que cette propriété est à domaine de valeurs fermé contrairement aux précédentes)

Import 4.1. - Texte XML-TEI avec un facsimile des pages de l'édition source et des illustrations dans le texte

Le document XML-TEI tdm80j.xml contient les références des images de l'édition source en ligne (attribut @facs de la balise <pb/>) et des illustrations dans le corps du texte (balises <graphic> avec l'attribut @url). Nous allons utiliser le module XTZ pour construire une édition synoptique

- l'interface du module permet d'indiquer les éléments « hors texte », « hors texte à éditer » et les « notes »
- l'option « construire l'édition "facs" » permet de construire l'édition synoptique
  - à partir des attributs @facs (URL absolu)
  - à partir d'un dossier d'images
- une feuille de style post-tokenisation permet d'ajouter des références par défaut pour les concordances

Import 5.1. - Transcription XML-TEI « multi-facette »

Nous allons partir du document .docx con tentant des styles de caractères (noms propres) et des raccourcis typographiques permettant de construire une édition à deux facettes : normalisée et diplomatique. Nous commencerons par une conversion vers XML-TEI avec Oxgarage et puis nous utiliserons le module d'import XTZ

- des feuilles de style 2-front et 3-posttok permettront de transformer les raccourcis typographiques en balises XML et de pré-tokeniser le document
- des feuilles de style 4-edition permettront de construire les deux facettes : normalisée (default) et diplomatique
- un sous-dossier contenant les images de colonnes
- un sous-dossier contenant des feuilles de style CSS

Discussion sur les projets et les corpus des participants

Pour chaque participant le souhaitant et en fonction du temps :

- Présentation du corpus et des objectifs de l'étude ;
- Étude du format du corpus ;
- Discussion sur la façon de l'encoder et de l'exploiter dans TXM ;
- Import dans TXM ;
- Exploitation avec TXM.

Annexe - éditeurs de texte populaires pour chaque système d'exploitation

Windows

- Notepad++ : http://notepad-plus.sourceforge.net (open-source)
- UltraEdit : http://www.ultraedit.com

- Emacs : http://savannah.gnu.org/projects/emacs (open-source, puissant mais complexe)
- Vi : http://www.vim.org (open-source, puissant mais complexe)

Mac OS X

- TextEdit : installé par défaut (open-source)
- Fraise : http://fr.wikipedia.org/wiki/Fraise_%28%C3%A9diteur_de_texte%29
- Aquamacs : http://aquamacs.org (open-source)

Linux

- Gedit : installé par défaut (open-source)
- SciTE : http://www.scintilla.org/SciTE.html (open-source)
- Emacs : http://savannah.gnu.org/projects/emacs (open-source, puissant mais complexe)
- Vi : http://www.vim.org (open-source, puissant mais complexe)

Voir d'autres possibilités dans la liste des éditeurs recensés dans le wiki de la TEI (en anglais) : http://wiki.tei-c.org/index.php/Editors