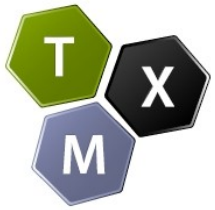


Atelier avancé TXM

- préparation de corpus
- import XTZ+CSV
- annotation

Alexei Lavrentev
UMR IHRIM, CNRS
textometrie@groupes.renater.fr

Diapositives réalisées par :
- Alexei Lavrentiev (AL)
- Bénédicte Pincemin (BP)
- Serge Heiden (SLH)

UMR5317 IHRIM, 2018

Programme de la séance

- Nouveautés de TXM 0.7.9
  - 14h00-14h30
- Module d'import XTZ + CSV
  - 14h30-15h30
- Annotation
  - 15h45-17h00
- Création et gestion de lexiques
  - 17h00-17h30

Qu'est-ce que la textométrie ?

Dans la lignée de la lexicométrie

- en particulier années 80 Saint-Cloud

Texto-métrie

- Analyse de données **textuelles** (en tenant compte de leur nature linguistique) : unités textuelles – unités lexicales
- Au moyen de **mesures** notamment statistiques : fréquence

Approche équilibrée

- **Quantitative** : décomptes, tris, statistiques
- **Qualitative** : moteur de recherche plein texte de motifs lexicaux et « retour au texte », on interprète en revenant aux formulations employées dans le texte et à partir de visualisations graphiques

Qu’est-ce que la textométrie ?

Interaction avec le corpus *versus*:

- *text mining*: extraire l’or à partir d’un tas de « minerai textuel »
  - pas de lien avec les données originales au moment de l’interprétation des résultats
  - choix de représentation sévères
  - vision non structurée du texte
- *information retrieval*: chercher le savoir dans un tas de données non (ou peu) structurées
  - contexte documentaire
  - pas de notion de déploiement textuel ni de structures textuelles

Logiciels de textométrie

généralistes

- Lexico 3, Hyperbase, Le Trameur, Weblex...

plus spécialisés

- Alceste, Iramuteq, DTM…

Points forts de TXM

- multiplateforme : windows, mac, linux, portail web
- la souplesse de l'import (formats et modules multiples)
- Unicode : nombreuses langues
- l'étiquetage à la volée (TreeTagger)
- une communauté d'utilisateurs ; extensions, macros et scripts pour personnalisation

TXM

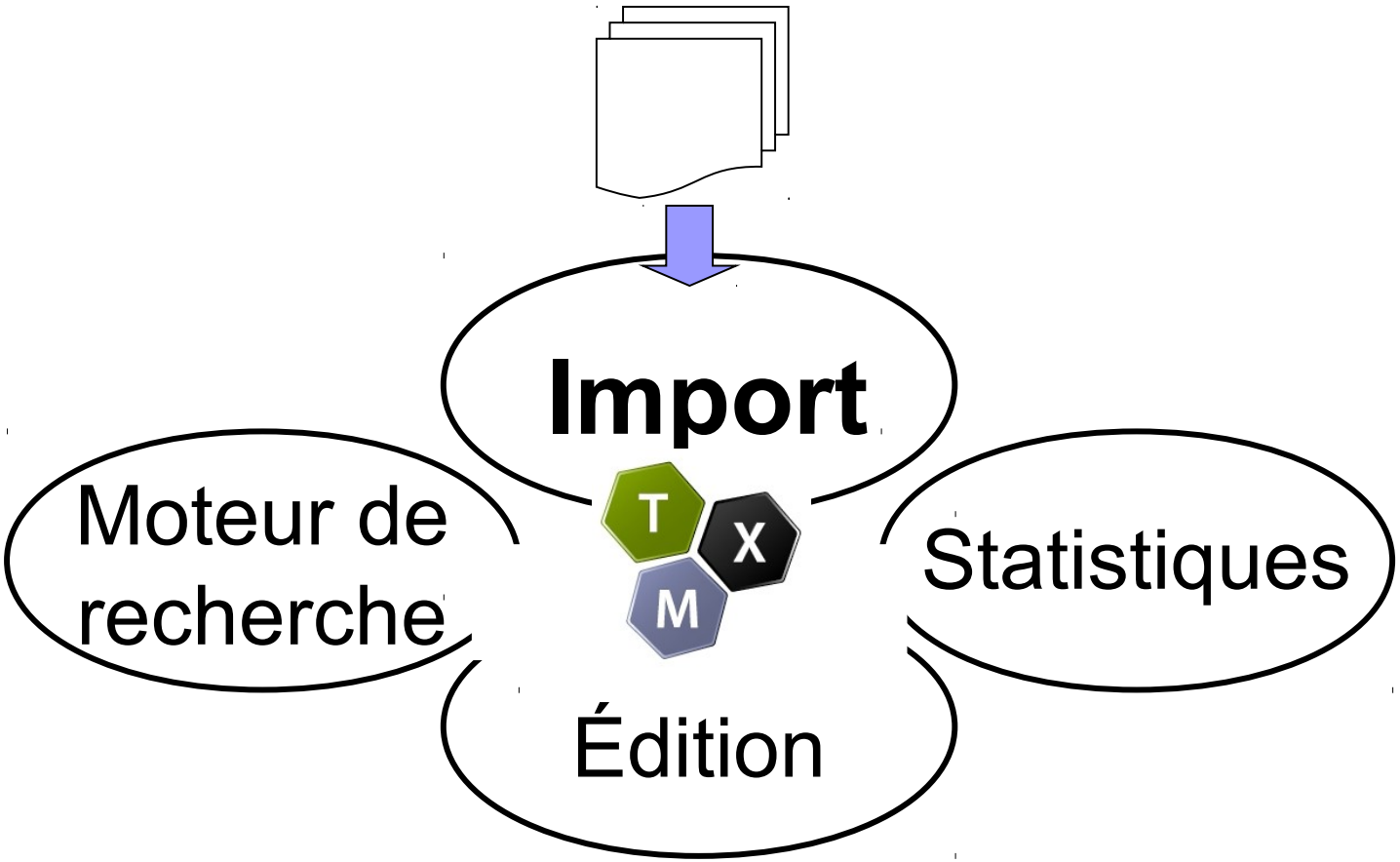
Développement

- Projet ANR Textométrie 2007-2010 (coord. S. Heiden)
- Depuis 2011 : Équipex Matrice, Labex Aslan, contributions de divers projets et laboratoires

Références web

- **Site projet : http://textometrie.org**
  - Téléchargement version pour poste (Linux/Windows/Mac)
  - Manuel utilisateur, documentation, bibliographie
- Portail de démonstration http://portal.textometrie.org/demo/
- Liste & wiki utilisateurs https://listes.cru.fr/sympa/info/txm-users
  - FAQ
  - Ateliers de formation (sur demande)
- Atelier initiation enregistré
  - http://txm.sourceforge.net/enregistrement_atelier_initiation_TXM_fr.html

Import dans TXM



3 familles de corpus

A) Corpus de **textes écrits** (TXT, XML, TEI) éditions alignées avec images de facsimilés

B) Corpus de **transcriptions d'enregistrements** (TRS), éventuellement *synchronisées* avec le son ou la vidéo

C) Corpus **multilingues alignés** (TMX), au niveau d'une structure textuelle comme la phrase ou le paragraphe

Modèle de corpus TXM

- **Unités textuelles** (livre, article, entretien…)
  Métadonnées (auteur, date, domaine, genre…)
  - **Structures internes** (phrase, paragraphe, sections…)
    Propriétés (numéro, titre...)
    - **Unités lexicales** (mots, mots composés)
      Propriétés (forme graphique, lemme, partie du discours…)
  - **Plans textuels**
    - Hors-texte (commentaires…)
    - Tours de parole, discours direct...
    - Langue principale (français…), Langue secondaire (latin…)
- Outils de TAL impliqués (lemmatiseurs…)
- Édition de texte pour le retour au texte
  - Pagination (sauts de pages)
  - Mise en page (styles), Média (Image, Audio, Vidéo)

Structure du « corpus binaire »

- archive zip (.txm)
  - import.xml : paramètres du corpus
  - data, registry : indexes du moteur CQP
  - txm : textes du corpus au format XML-TEI TXM (facultatif)
  - HTML : pages de l'édition / des éditions
    - default
      - css
    - facs
  - ...

Modules d'import TXM

formats de corpus en entrée

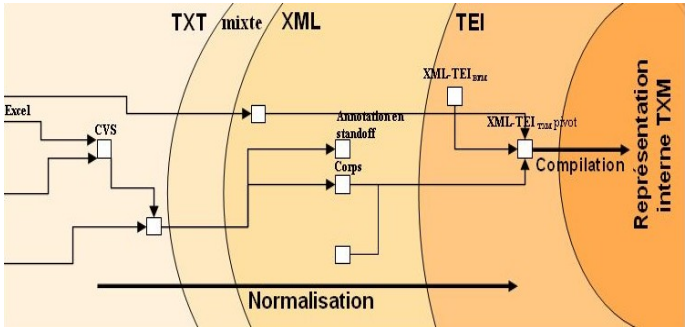
- Formats propriétaires divers : Hyperbase, Alceste, CNR (Cordial)
- *Calibre* – ePub

- Copier/Coller
- TXT Unicode+CSV (metadata) : dossier de fichiers texte brut
- XML/w+CSV : dossier de fichiers XML
- (TXM 0.7.8) **XML-XTZ** : reconfiguration des textes, éditions stylées

- XML-TEI P5 **BFM** : personnalisation de la TEI
- XML-TEI P5 **FRANTEXT (textes)**
- XML-TEI P5 **FRANTEXT (résultats de requêtes)**
- XML-TEI-TXM : XML compatible TEI+TAL (pivot)

- XML-Transcriber+CSV – transcriptions audio alignées
- *XML-TMX* – corpus multilingues alignés
- *XML-PPS*-Factiva – portail de presse

Processus d'import et d'analyse



- répertoire de fichiers TXT
- textes XML + metadonnées
- textes annotés TAL
- textes XML-TXM TEI
- Contrastes : sous-corpus & partition
- Structures
- Facettes lexicales

TreeTagger

TXM

Métadonnées

- metadata.csv
  - format personnalisable (préférences TXM)
    - par défaut : UTF-8 / virgule / "
  - une colonne « id » = nom du fichier
  - une colonne « textorder » (facultative)
- attributs de la balise <text>
  - éviter les majuscules dans les noms d'attribut
- en-tête TEI
  - utiliser une feuille XSLT pour la projection vers les attributs de <text>

Module d'import XTZ+CSV

- Paramètres d'import
  - Éditions
    - pagination : nombre de mots et/ou balise
    - édition « facs » (synoptique)
  - Plans textuels
    - hors texte → supprimé du corpus
    - hors texte à éditer → non indexé, mais peut s'afficher
    - note → non indexé, s'affiche en note de bas de page
    - (éléments milestones)

Module d'import XTZ+CSV

- Balises TEI interprétées
  - voir Manuel TXM, section 6.1.5
  - <text> (1 par fichier)
    - → ne pas utiliser teiCorpus, group
  - <w> (peut être modifié dans le formulaire)
  - pour les éditions : pb, head, p, hi, emph, list, table…
  - les autres balises (TEI ou pas) deviennent des structures

Module d'import XTZ+CSV

- Personnalisation des éditions avec CSS
  - sous-dossier « css »
    - [NOMDUCORPUS].css
    - TXM.css
  - polices d'affichage, mise en page, couleurs, etc.
    - titres, notes, mises en relief (<hi>, <emph>)
  - toutes les balises du document d'origine ne sont pas disponibles

Langage XSLT

- Conçu pour le traitement d'XML
  - 3 versions : TXM prend en charge XSLT 2
    - expressions régulières, tokenisation, fusion / division de documents
- Un script XSLT est
  - un document XML
  - une série de *templates* recherchées et appliquées dans le document traité
- Xpath : navigation dans l'arbre XML

Module d'import XTZ+CSV

- Transformations XSLT
  - 1-split-merge : reconfiguration des fichiers
    - actuellement bogué (TXM 0.7.9)
  - 2-front : préparation à la tokenisation
  - 3-posttok :
    - réglage de tokenisation,
    - création de références pour les concordances,
    - projection de propriétés de mots
  - 4-edition :
    - personnalisation de l'édition par défaut,
    - création d'éditions supplémentaires

Module d'import XTZ+CSV

- Bibliothèque XSLT de TXM
  - 1-split-merge : txm-split-xces-ids-corpus2text.xsl
    - transforme un fichier unique d'un corpus XCES-IDS en autant de fichiers que de textes
    - étape boguée sous TXM 0.7.9, la feuille peut être utilisée avec la macro ExecXSL
  - 2-front :txm-front-idsHeader2textAtt.xsl
    - projette les métadonnées de l'entête IDS (XCES) vers les attributs de la balise <text>

Module d'import XTZ+CSV

- Bibliothèque XSLT de TXM
  - 3-posttok :
    - txm-posttokAddRef.xsl
      - ajoute des références pour les concordances
    - txm-posttok-structure2wordAtt.xsl
      - crée un attribut pour indiquer la profondeur d'imbrication dans une structure (ex. <q>)
    - txm-posttok-unBreakWords.xsl
      - recolle les mots coupés par un saut de ligne

Module d'import XTZ+CSV

- Bibliothèque XSLT de TXM
  - 4-edition :
    - 1-default-html.xsl
      - prépare un fichier HTML pour l'édition par défaut
      - toutes les balises deviennent ou avec l'attribut @class → pratique pour le stylage CSS
    - 2-default-pager.xsl
      - découpe le fichier HTML en pages
    - On peut ajouter d'autres éditions
      - en incrémentant les chiffres pour l'ordre d'application
      - en remplaçant « default » par le nom de la nouvelle édition (ex. 3-dipl-html.xsl et 4-dipl-pager.xsl)

Annotation dans TXM

- Annotation simple par concordance
  - projet BHE (UMR Larhra & IHRIM)
  - disponible pour les corpus importés avec le module XTZ+CSV
  - bouton « crayon » dans les concordance
  - ajout d'une categorie ou d'une paire catégorie + valeur
    - → création d'une structure
  - sauvegarde par simple clic

Annotation dans TXM

- Annotation simple par concordance
  - exploitation
    - <span_ref="persName"> [] expand to span
  - limitations
    - ne concerne pas les propriétés de mots
    - chevauchement interdit (avec <p> par exemple)
  - Pour plus de détail
    - voir le chapitre 9 du Manuel TXM

Annotation dans TXM

- Correction et ajout de propriétés de mots
  - TXM 0.8 (disponible avant fin 2018)
    - par concordance
  - TXM 0.7.9 :
    - macro « annotation / BuildWordPropTable »
    - annotation en concordance dans un logiciel externe (Calc, Excel…)
    - macro « annotation / InjectWordPropTable »
    - réimport du corpus (XML-TEI TXM)
    - Tutoriel
    - https://groupes.renater.fr/wiki/txm-users/public/tutoriel_correction_mots

Annotation dans TXM

- Annotation avec un modèle Unité-Relation-Schéma (URS) au fil du texte
  - projet ANR Democrat (resp. F. Landragin)
  - reproduction
  - extension Analec

Annotation dans TXM

- Extension « TreeTagger »
  - projet ANR-DFG PaLaFra
    - resp. C. Guillot-Barbance (ENS Lyon) et M. Selig (U. Regensburg)
  - installer une « extension tierce » à partir de
    - http://textometrie.ens-lyon.fr/dist/palafra

Annotation dans TXM

- Extension « TreeTagger »
  - création d'un modèle de langue
    - pour créer un nouveau modèle de langue
    - « corpus gold » (annoté) (~ 100 000 mots minimum)
    - commande « TreeTagger > Train »
  - application d'un modèle de langue sur un corpus TXM
    - commande « TreeTagger > Apply »

Télécharger le dossier de travail

http://bit.ly/2Iz0yr7