Préparer un corpus pour TXM

Eva Schaeffer-Lacroix

Préparer un corpus pour TXM

Auteure : Eva Schaeffer-Lacroix, maître de conférences (section 12), Laboratoire CeLiSo (Centre de Linguistique en Sorbonne), Université Paris-Sorbonne.

Carnet Hypothèses accompagnant cette formation : https://initxm.hypotheses.org/

Conseils avant de commencer

S'il s'agit de créer un corpus composé d'un nombre assez important de parties, il est prudent de faire d'abord un corpus d'essai avec trois ou quatre fichiers avant de passer à la mise en corpus des données dans leur intégralité. Cela permet de se faire la main et de voir si l'on a fait le bon choix avec les intitulés de certains éléments (nom du corpus, catégories, etc.).

Il est également utile de créer un dossier de secours dans lequel figurent la version d'origine des fichiers et du fichier des métadonnées[2].

Checklist

❑ 1. Créer un dossier portant le nom du corpus tel qu'il doit apparaître sur TXM. Exemple : ECONOMIE.

❑ 2. Y déposer les fichiers txt (texte brut) qui composent les différentes parties du corpus (par exemple, les articles correspondant à l'une des disciplines universitaires représentées dans le *corpus ChambersLeBaron* (Chambers & Le Baron 2006) qui illustre les propos de ce tutoriel.

❑ 3. Nommer les fichiers selon un procédé cohérent, reconnaissable et informatif. Il peut s'agir d'un mot représentatif (le mot principal du titre d'une publication, par exemple). Parfois, on trouve des codes qui renseignent sur des caractéristiques essentielles du corpus comme ceux utilisés dans le *corpus ChambersLeBaron* : <EC-EI-AB-XX-96-03>.

Décryptage de ce code : il s'agit d'un texte de la discipline "**EC**onomie" qui a paru dans le journal "**E**conomie **i**nternationale" et dont l'auteure s'appelle **A**urélie **B**oubel. Suivent deux indications que je n'arrive pas à décrypter, puis l'année de publication : 20**03**.

❑ 4. Normaliser les fichiers texte : guillemets simples et doubles multiformes à remplacer par des guillemets droits ; enlever les métadonnées dans des fichiers que l'on ne souhaite pas structurer en

---

[1] "BY SA" veut dire :
"Attribution — Vous devez créditer l'Œuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'Oeuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son Oeuvre.

Partage dans les Mêmes Conditions — Dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l'Oeuvre originale, vous devez diffuser l'Oeuvre modifiée dans les même conditions, c'est à dire avec la même licence avec laquelle l'Oeuvre originale a été diffusée."

(Source : https://creativecommons.org/licenses/by-sa/3.0/fr/)

[2] "Une métadonnée est un ensemble structuré d'informations décrivant une ressource quelconque" ; elle peut être descriptive (auteur·e, titre, genre textuel, année de publication, etc.), de structure (pour renseigner sur les différents documents dont se compose une ressource) ou administrative (par exemple, pour donner des informations sur les droits d'utilisation). Source : BNF - Document numérique et métadonnées.

interne. Il peut, entre autres, s'agir des textes explicatifs/législatifs qui figurent au début et à la fin d'un document publié sur *Gutenberg Project*.

❑ 5. Enregistrer les textes (au moins ceux écrits en une autre langue que l'anglais) sous UTF-8.

❑ 6. Créer un fichier Excel contenant les métadonnées[3] et le nommer <metadata> (et pas autrement).

❑ 7. Créer une première ligne réservée aux catégories.

La première colonne de la première ligne sera obligatoirement nommée <id> (identifiant). Elle contient les noms des fichiers qui doivent correspondre exactement aux noms des fichiers texte ; elle contiendra un seul mot. Dans la section FAQ de *Textométrie*, les contraintes sont précisées comme suit :

> Les intitulés des colonnes (qui deviendront les noms des propriétés associées aux textes) doivent respecter les contraintes suivantes (liées au langage CQL) :
>
> - pas d'espace,
> - pas de caractères accentués,
> - pas de majuscule,
> - pas de chiffre en premier ni en dernier caractère,
> - d'une manière générale, éviter les ponctuations et les caractères non alphanumériques.
>
> Le fichier metadata.csv doit être placé à coté des fichiers source, dans le même répertoire.

Les catégories autres que <id> peuvent être choisies en fonction de ce qui intéresse l'auteur·e du corpus. En l'occurrence, <discipline>, <auteur>, <titre> et <journal> peuvent convenir. On évitera des entrées trop longues ; éventuellement, on ne mettra que le début des titres des articles cités en tableau 1.

id	discipline	auteur	titre	journal
EC-EI-AB-XX-96-03	Economics	Boubel Aurélie	LES INVESTISSEURS INSTITUTIONNELS ET L'ÉPARGNE RETRAITE	Économie internationale
EC-EI-DL-XX-86-01	Economics	Labaronne Daniel. Siroën Jean-Marc	PRIVATISATION ET CROISSANCE DANS LES PAYS DE L'EST	

Tableau 1 – Créer un fichier contenant les métadonnées.

❑ 8. Créer une ligne par partie du corpus. Le corpus *ChambersLeBaron* (discipline *Economics*) contient 16 fichiers ; on aura donc 17 lignes, celle des catégories et les 16 lignes des différents fichiers contenant chacun un article de journal.

Une remarque : dans le tableau 1, les titres sont écrits en majuscules. Ce n'est pas obligatoire ; cela correspond à la présentation dans les articles (j'ai juste copié-collé les titres dans les fichiers texte).

La deuxième entrée (id : EC-EI-DL-XX-86-01) concerne un article écrit par deux auteur·es. J'ai choisi de séparer les noms par un point.

❑ 9. Enregistrer le fichier des métadonnées dans un dossier de secours, en le renommant de façon explicite, par exemple, <metada_V0> ce qui correspond à "version d'origine".

❑ 10. Convertir le fichier <metadata> qui est dans le dossier du corpus en format <csv UTF-8 (délimité par des virgules)>.

❑ 11. Ouvrir le fichier metadata avec *LibreOfficeCalcPortable* pour vérifier si le formatage est correct (encodage des caractères en UTF-8, choix de langue correspondant à celle du corpus) et si la délimitation a

[3] Il est également intéressant de lire la section FAQ du site Textométrie au sujet des métadonnées.

bien été effectuée par des virgules (et non par des points-virgules). C'est là aussi que l'on voit si les caractères spéciaux (accents français, ß allemand, etc.) sont encodés correctement. Procéder aux rectifications si nécessaire et réenregistrer dans le dossier du corpus.

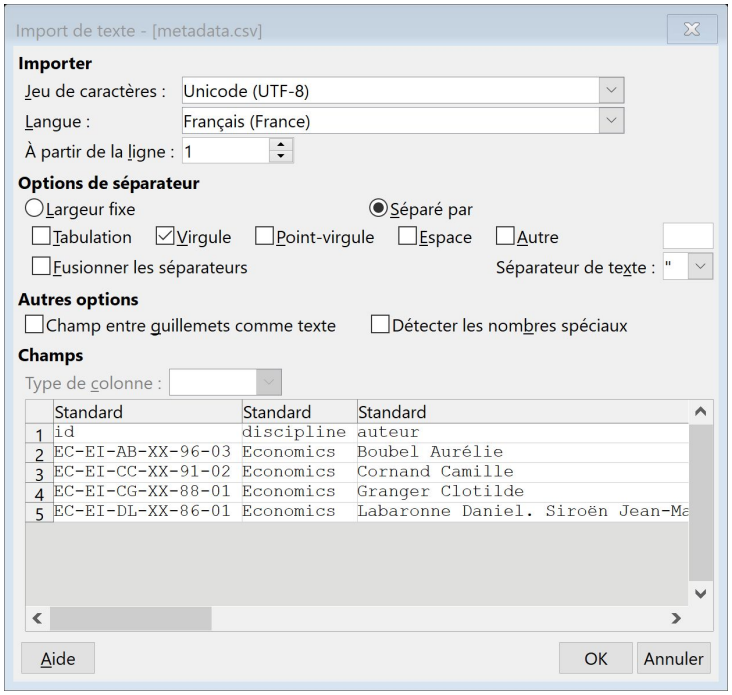

Figure 1 – Vérification de l'encodage, de la langue et du séparateur.

Importer un corpus dans TXM

Import simple

Si l'on n'a pas l'intention de travailler avec des corpus partitionnés et/ou des sous-corpus, on peut opter pour l'import par copier-coller. La commande <Fichier – Importer – Presse papier> transformera tout fichier texte copié dans le presse-papier (à l'aide de la commande <sélectionner tout – contrôle c>) en un corpus simple et non structuré (sauf si vous l'avez structuré en interne, à l'aide de balises XML, par exemple). Un inconvénient : si l'on choisit ce procédé, le corpus s'appellera "Pressepapier1" ou "Pressepapier2", etc.

Pour des textes en une langue autre que le français, il faudra procéder à un paramétrage supplémentaire : commande <Outils – Préférences – TXM – Utilisateur – Import> (ou tapez "Import" dans la case tout en haut à gauche pour sauter les étapes "TXM" et "Utilisateur).

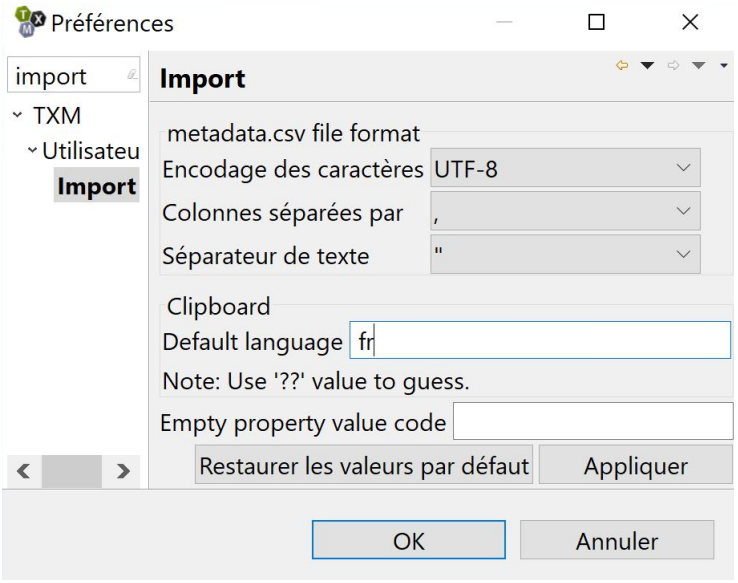

Figure 3 – Paramétrage de l'encodage des caractères et de la langue.

Choisir dans le menu déroulant l'encodage UTF-8 et remplacer <fr> par <de> pour des textes en allemand, <en> pour l'anglais, <ru> pour le russe, etc.

Import de fichiers avec métadonnées

Si l'on dispose d'un fichier contenant les métadonnées, on peut opter pour un import permettant de séparer par la suite le corpus en autant de parties que de fichiers (on créer une "partition") ou d'en combiner deux ou plus (on crée des "sous-corpus").

La commande <Fichier – Importer – TXT + CVS> permettra d'accéder à un menu à partir duquel on indiquera à TXM l'endroit où se trouver le dossier corpus dans l'ordinateur : <Sélectionner le répertoire des sources>.

Il faudra vérifier l'encodage des caractères et la langue principale qui, par défaut, est réglé sur <fr> (textes en langue française).

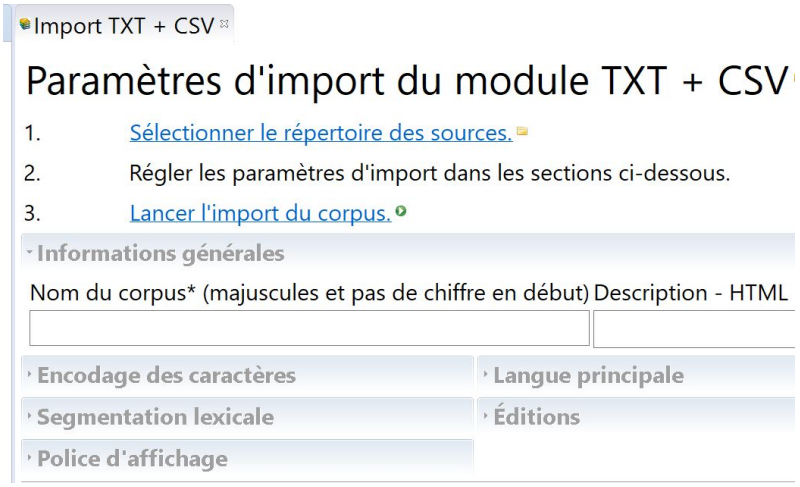

Figure 2 – Import du module TXT + CSV.

Il est également possible de vérifier le bon affichage des métadonnées en cliquant sur <Aperçu des métadonnées> en bas du module d'import. Si les caractères ne sont pas encodés correctement, il faudra vérifier si UTF-8 a été sélectionné partout, y compris dans le ménu "Options – Préférences – (...) Import".

On lancera ensuite l'import du corpus. Son nom apparaîtra alors dans le menu à gauche.

On peut ensuite cliquer sur <Description> pour voir son contenu (nombre de mots, noms des fichiers, etc.). Le menu <Édition> permet d'afficher le texte suivi et de vérifier si les données ont bien été annotées à l'aide de *TreeTagger*.

Pour que le réglage de la langue fasse effet, il faut avoir installé *TreeTagger* et les modèles de langue correspondants et paramétré les chemins dans TXM (v. page d'aide dans TXM : menu Aide > Installer TreeTagger ; v. tutoriel filmé) - d'où l'importance de placer le dossier d'installation de *TreeTagger* à un endroit facilement repérable de son ordinateur.

Références

Chambers, Angela et Le Baron, Florence (2006). *Corpus Chambers-Le Baron d'articles de recherche en français*. http://ota.ahds.ac.uk/headers/2527.xml

Heiden, Serge, Magué, Jean-Philippe et Pincemin, Bénédicte (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », dans S. Bolasco (dir.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data.* JADT 2010. Vol. 2, 1021-1032. Rome : Edizioni Universitarie di Lettere Economia Diritto. https://halshs.archives-ouvertes.fr/halshs-00549779/fr/

Textométrie (nd). http://textometrie.ens-lyon.fr/

- Rubrique Documentation TXM.
- Manuel de TXM 0.7 : pour imprimer (PDF) ; pour lire en ligne (HTML).

Le Wiki de la liste TXM users (nd). https://groupes.renater.fr/wiki/txm-users/

- Section FAQ