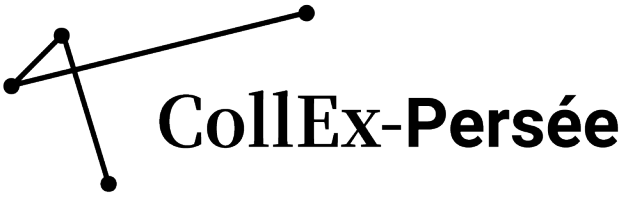
Reste à faire sur le traitement de La Grande Encyclopédie

A. BRENON D. VIGIER

11 mars 2020



Le texte ci-dessous propose un bilan d'étape du travail conduit actuellement sur le texte de *La Grande Encyclopédie* (LGE) au sein du projet CollEx-Persée LGE.

Bilan

Le projet Collex Persée a débuté il y a un peu plus de sept mois. Ce temps a surtout été investi sur les étapes préparatoires nécessaires au traitement des volumes de La Grande Encyclopédie (LGE), en amont de l'étiquetage morpho-syntaxique.

Normalisation

Le projet vise à enrichir le texte des 31 volumes de LGE en l'annotant morpho-syntaxiquement, pour nos propres besoins d'analyse linguistique (ICAR) mais également pour que la communauté puisse réutiliser les textes pour d'autres projets.

Nous avons dû commencer par définir un encodage standard assez flexible pour représenter les particularités de LGE, notamment les longs développements très structurés (présence de titres, de sous-titres et sections numérotées etc.) sur plusieurs dizaines de pages (voir par ex. l'article "Administration" sur Gallica, p.581 & sq.).

Après de nombreuses discussions internes (BNF, LICORN, INRIA, ...) et externes au projet, notamment sur la liste de diffusion *TEI-L*, nous avons considéré que certaines des structures propres au module *dictionaries* de la XML-TEI étaient difficilement compatibles avec la structure de l'article encyclopédique. Par exemple, la présence attendue d'une définition unique en début d'article, clairement délimitée et facilement indentifiable, ou encore la structure profonde en parties et sous-parties de certains articles. toutes ces caractéristiques ne permettraient pas d'utiliser les balises spécifiques à ce module[1].

Nous avons donc défini et proposé un nouveau schéma d'encodage parfaitement conforme à la XML-TEI tout en n'optant pas pour le module *dictionaries*, et permettant de rendre compte des particularité de LGE.

Traitements

Nous avons reçu très tôt les premiers tomes numérisés et OCRisés de la BNF, ce qui nous a permis d'entamer les premières tâches de nettoyage nécessaires pour traiter les fichiers avec l'outil GROBID, et d'estimer la complexité pour généraliser le processus et pouvoir l'appliquer efficacement à l'ensemble des 31 tomes qui constituent LGE.

Nous avons réussi à entraîner avec succès les deux premières couches de la pile d'analyseur *Conditional Random Fields* (CRF) qui constitue GROBID. Il s'agit des couches dénommées `DictionarySegmentation` et `DictionaryBodySegmentation`, visant respectivement à séparer le corps du texte des éléments péritextes et les articles entre eux à l'intérieur du corps du texte.

Pour la deuxième couche, ce chantier s'est accompagné d'un petit travail de reconfiguration de GROBID puisqu'utilisant les balises du module TEI *dictionaries* par défaut, il encode les articles séparés à l'issue de l'analyse de la deuxième couche dans des éléments `<entry/>` là où nous avons fait le choix d'utiliser des éléments `<div/>`. Ce travail a été conduit avec succès et nous a permis de produire 3 séquences continues de 5 pages, 5 pages et 6 pages choisies aléatoirement dans le premier tome contenant les articles clairement identifiés dans des `<div/>` distincts.

Points d'étapes

Pour atteindre l'état final que nous souhaitons, le corpus doit passer par 4 étapes distinctes entre lesquelles il est aisé de décrire les transformations nécessaires.

A) L'encyclopédie numérisée, OCRisée et nettoyée, un fichier PDF par tome
B) Le texte encodé de manière automatique par le logiciel GROBID en XML-TEI, un fichier par tome

[1] vous trouverez à ce sujet, en annexe de ce bilan, le document que nous avons projeté lors de notre dernière réunion (let 15/11/2019) à la BNF à laquelle ont participé S. Cretin, C. Jacquet-Pfau, Y. Le Guillou, J.P. Moreux, G. Willams et nous-mêmes.

C) Les tomes séparés en autant de fichiers (toujours encodés en XML-TEI) que d'articles puisque l'unité de base d'étude de notre corpus est l'article. Nous voulons en effet pouvoir étudier spécifiquement le discours dans les articles d'un domaine ou d'un auteur particulier.
D) Étape finale, le texte traité par la chaîne de traitement PRESTO, sous la forme de fichiers XML-TEI importables dans TXM, un par article toujours

Éléments de la chaîne à concevoir

A) PDF numérisés et OCRisés

Tous les PDF ont été livrés par la BNF et sont disponibles sous forme de fichiers PDF ou de répertoires contenant un fichier XML-ALTO par page. Il reste à nettoyer les fichiers PDF, seuls acceptés par GROBID, ce qui requiert

- **A.1** d'écrire un algorithme pour écrémer le texte des scories restant de l'OCRisation *ou bien*
- **A.1'** une trentaine de jours pour traiter manuellement les tomes
- **A.2** il existe également des scories de position, avec des éléments péritextes inclus au milieu du flot des pages, il faudrait idéalement pouvoir automatiquement corriger la position de ces éléments ou bien trouver une solution pour les filtrer en sortie
- **A.3** de finir le code de la librairie Hufflepdf pour pouvoir éditer les PDFs et appliquer les algorithmes précédents

B) GROBID

Nous avons pour l'instant suite à l'entraînement de GROBID obtenu un module permettant d'encoder les tomes jusqu'au niveau des balises `<div/>` autour de chaque article, mais pas encore davantage de finesse (pour ce qui concerne les vedettes des articles, les paragraphes du corps de l'article etc.). Voici les tâches restantes pour parvenir pleinement au point B défini plus haut.

- **B.1** il reste à entraîner 3 autres modèles d'analyse de GROBID, `LexicalEntry` puis `Form` et `Sense` (qui correspondent à la même couche, procédant respectivement à l'analyse de la vedette et du corps de l'article), et peut-être une 4ème d'après le code source de GROBID (nommée `SubSense` et qui permettrait de structure plus finement le corps de l'article); comme expliqué ci-dessus pour la couche `DictionaryBodySegmentation`, il s'agit également à chaque modèle de changer légèrement le code source de GROBID pour produire en sortie les balises non-*dictionaries* que nous avons dû utiliser pour définir notre schéma d'encodage pour les articles
- **B.2** il faut aussi tester la faisabilité de l'annotation d'un volume entier avec l'API, ce qui n'est pas documenté, toutes les annotations ayant pour l'instant été faites dans l'interface web distribuée avec le logiciel GROBID. Il s'agit d'essayer de parler directement avec l'API mise à disposition par GROBID via l'outil `curl`, de voir les limites de taille et d'essayer d'allouer davantage de mémoire le cas échéant

- **B.3** le degré de finesse dans l'annotation pouvant être obtenu avec GROBID étant difficile à anticiper, notamment en ce qui concerne les subdivisions en parties et sous-parties, il pourrait s'avérer nécessaire de chercher à adapter un modèle CRF existant dans le logiciel pour tâcher d'en faire une nouvelle couche d'analyse capable d'annoter correctement la structure profonde des articles

B') Raccourci / Plan de secours ?

En cas de réelle impossibilité, trouver un moyen de faire l'OLR et la conversion vers XML-TEI à partir de l'ALTO (**B'** remplace toute la section GROBID ainsi que le code d'édition de PDF puisque le texte étant en clair dans l'ALTO, sa modification est «gratuite»)

C) Découpage

Un script existe, assez général, qui permet de découper un fichier XML suivant un `XPath` donné (il produit autant de fichier que de nœuds désignés, contenant chacun toute l'arborescence au-dessus de ce nœud et ce nœud seul débarrassé des nœuds adelphes).

- **C.1** La librairie avec laquelle cet outil est écrit considère les tabulations comme des caractères spéciaux; proposer un correctif à la librairie ou tout simplement filtrer la sortie
- **C.2** Dans notre encyclopédie, des balises `<fw/>` contenant les éléments péri-textes de la page (numéro de page, intervalle des mots compris dans la page...) peuvent se trouver hors des `<div/>` des articles, ce qui complexifie le découpage. Il faut raffiner la manière de le faire (en l'état actuel, le `<fw/>` se retrouve dans tous les fichiers correspondant aux `<div/>` qui le suivent), et soit le garder seulement dans le fichier de l'article correspondant à la première `<div/>` immédiatement après, soit l'enlever complètement de tous les fichiers, ce qui bien sûr perdrait de l'information, rendant nos fichiers moins utiles pour des projets ultérieurs qui s'intéresseraient justement aux éléments péri-textes, mais qui n'aurait pas d'impact immédiat pour notre projet s'intéressant au discours encyclopédique seulement.

D) PRESTO

Pour l'instant, la chaîne de traitement presto réduit encore les fichiers en un format «plat», une liste de balises `<w/>`. L'import dans TXM d'un fichier avec une structure de balises plus complète a été testé avec succès mais il reste à trouver un moyen pour que la chaîne de traitement en amont ne détruise pas ces balises, mais les garde en plus de leur ajouter des annotations morpho-syntaxiques.

- **D.1** Une solution consiste à extraire le texte brut du XML en notant les positions (dans le texte brut) auxquelles les balises s'ouvrent et se ferment, à annoter le texte brut, à faire subir le même sort d'analyse des positions aux balises trouvées par l'étiqueteur, à fusionner les références de positions

de balise puis à régénerer le texte contenant les deux jeux de balises. Il a l'air de s'agir d'une approche similaire à celle employée par bgaiffe dans son XMLMixer, il pourrait être intéressant de voir si son code peut être réemployé tel quel ou s'il est au moins possible de s'en inspirer pour écrire un outil appliquant cet algorithme.

Méta

Les parties précédentes détaillent les tâches restant à accomplir pour passer d'étape en étape, mais ce que nous cherchons à produire est un outil complet qui permette le traitement de bout en bout.

- **ABCD.1** Écrire réellement une telle chaîne, sous forme d'un script qui exploite tous les logiciels écrits pour réaliser les étapes précédentes et permette en une opération d'obtenir des fichiers XML-TEI annotés morpho-syntaxiquement pour chaque article à partir d'un tome PDF en entrée.