II Journée Internationale :
« Anciens textes, nouveaux outils :
des corpus alignés aux éditions multiples »
19 octobre 2018. Colegio de España (Paris)

Depuis maintenant une vingtaine d'années, la création d'un domaine où se croisent les humanités et les technologies informatiques, celui des humanités numériques, a vu apparaître de nouvelles techniques et de nouvelles pratiques dans le domaine de l'édition (éditions numériques, éditions parallèles), de la diffusion (bases de données, plateformes en ligne, réseaux collaboratifs, programmes de visualisation des données), de la connaissance des textes (moteurs de recherche, logiciels de calcul lexicométrique, lemmatisseurs, concordanciers...).

Ces innovations techniques ont considérablement modifié l'approche et même l'objet d'étude des chercheurs travaillant sur les textes anciens.

Malgré les succès de ces études, les chercheurs en sciences humaines peuvent parfois rencontrer des difficultés pour intégrer le domaine des humanités numériques car cela exige la maîtrise de connaissances techniques qui ne sont pas toujours facilement accessibles. La collaboration avec des spécialistes en informatique et en intelligence artificielle s'avère ainsi essentielle pour le développement de cette branche d'étude.

Après la célébration d'une première journée internationale en octobre 2015 intitulée «Anciens textes, nouveaux outils: la philologie à l'ère numérique», nous organisons en octobre 2018 une deuxième journée, coïncidant avec la création dans notre laboratoire de recherche (Laboratoire d'Études Romanes, EA 4385) de l'atelier «Romanités Numériques» (prochain quinquennat). Cette deuxième journée poursuivra la réflexion entamée lors de la première rencontre, en explorant cette fois les possibilités qu'offrent les technologies numériques pour l'étude et l'édition des textes anciens à travers les corpus alignés et les éditions multiples dans le but de créer à terme une plateforme d'édition multiple et d'étude des textes en langues romanes. Nous souhaitons en 2018 réfléchir aux possibilités que ces plateformes offrent pour la connaissance des textes et pour l'étude de la variation linguistique (aussi bien en diachronie qu'en synchronie) dans les langues romanes.

La journée se veut ouverte à un public large d'étudiants et de chercheurs intéressés par les humanités numériques, ainsi qu'à des spécialistes des langues romanes.

UNIVERSITÉ PARIS 8
VINCENNES-SAINT-DENIS

Journée d'études

Vendredi 19 octobre 2019

Organisation:
Atelier « Romanités Numériques »
Laboratoire d'Études Romanes (EA 4365). Université Paris 8

II Journée Internationale

Anciens textes, nouveaux outils : des corpus alignés aux éditions multiples

Colegio de España

COLEGIO DE ESPAÑA
7 E Bd. Jourdan, 75014 PARIS

Contact :
Marta López Izquierdo (marta.li@univ-paris8.fr)

09h15 **Marta López Izquierdo** *(Université Paris 8)*

▸ **Présentation de la journée et du projet MULTIPLED**

Nous présenterons notre projet de construction d'une plateforme pour l'édition alignée de différentes versions d'un même texte ou d'un texte et ses traductions dans le cadre de l'Atelier « Romanités Numériques » de l'Université Paris 8 (Laboratoire d'Études Romanes, EA 4385).

09h30 **Shalev Vayness** *(ISAKO)*

▸ **Tout ce que vous avez toujours voulu savoir sur le XML sans jamais oser le demander**

But de l'intervention:
Expliquer et démystifier les notions et principaux éléments de l'informatique éditoriale et des bibliothèques numériques applicables au projet *Humanités Numérique*.
Description :
- Notions, termes et éléments principaux de l'informatique éditoriale et des bibliothèques numériques.
- Ce qu'est le langage XML et pourquoi il est si utile et si utilisé. Les différents types de XML ; distinguer XML et HTML ; le XML-TEI. - Cas concrets : de la préparation des données aux fonctionnalités en ligne. Exemples présentés : la bibliothèque numérique Gallica (BnF) ; les archives de la revue ESPRIT ; les archives du journal italien *Corriere della Serra.* - Méthodes et algorithmes fréquemment utilisés dans les étapes suivantes : indexation, balisage, recherche, comparaison. - Quelques clés pratiques pour aider à définir (et à expliquer) les besoins informatiques au sein d'un projet de recherche en lettres.

10h00 **Alejandro Bía Platas** *(Universidad Miguel Hernández)*

▸ **Alineación automática y otras herramientas informáticas aplicables a las Humanidades Digitales**

En esta ponencia, hablaremos de la alineación de textos automática o asistida por ordenador. Veremos los antecedentes de esta tecnología, desde Gale y Church (1993) en adelante, incluyendo Bilingual Sentence Aligner (Moore (2002), Hunalign (Varga et al. (2005), Gargantua (Braune and Fraser (2010), Bleualign (Sennrich and Volk (2010) y otros. Hablaremos también de proyectos propios como MDR, Match-Detect-Reveal (Zaslavsky, Bia, Monostory, 2001) y TRACE Aligner (2012-2016), desarrollado dentro del proyecto TRACE (2012-2017). TRACE es sinónimo de "traducción y censura", y consiste en el estudio de cómo la traducción se ve afectada por la censura en España desde 1939 a 1985, utilizando para el estudio diferentes géneros textuales y diferentes combinaciones de lenguas. TRACE realiza un uso intensivo de técnicas informáticas, como el marcado XML, la codificación TEI, las memorias de traducción (TMX), con el fin de construir y explotar corpus paralelos alineados multilingües.
Hablaremos del uso de TMX (Translation Memory eXchange) y de Bitext2tmx (Antón et al., 2008), un programa para alinear y segmentar oraciones traducidas, contenidas en dos archivos de texto sin formato, y generar una memoria de traducción (formato TMX). Comentaremos también las posibilidades del TEI (Text Encoding Initiative) para la creación de documentos TEI plurilingües.

11h00 **Stéphane Patin** *(Université Paris Diderot)*

▸ **Cuestiones de direccionalidad en los corpus paralelos: ejemplo del *Europarl***

Europarl es un corpus paralelo multilingüe constituido por las intervenciones de los diputados europeos traducidas en las 24 lenguas oficiales por la Dirección General de Traducción de la Union Europea. Dicho corpus tiene dos versiones distintas, una, multidireccional, en la cual, la lengua fuente no es forzosamente la lengua original, y otra, bidireccional, donde los textos de la lengua fuente, original, están traducidos en una lengua meta original, y viceversa. Ahora bien, según ese criterio de direccionalidad, se pueden observar, a veces, unas traducciones distintas.

11h45 **Jean-Gabriel Ganascia** *(Université Paris Sorbonne)*

▸ **Versant littéraire des humanités numériques**
- alignement d'états de textes et étude de l'intertextualité

Après une présentation générale du versant littéraire des humanités numériques et de notre approche épistémologique de ce domaine, nous ferons état des travaux que nous poursuivons depuis une quinzaine d'années en collaboration avec différents laboratoires sur l'alignement d'états de texte et sur l'étude de l'intertextualité. Il s'agira, dans un premier temps, de présenter les résultats obtenus en décrivant différents logiciels que nous avons réalisés, dont MEDITE pour l'alignement d'états de textes, Phœbus pour la détection de similarités et Galaxies pour la visualisation de communautés de similarités. Ensuite, nous ferons état de l'histoire de ces différents logiciels, des raisons qui nous ont poussé à les concevoir, des résultats obtenus et des collaborations que nous avons eu avec les équipes littéraires, entre autre avec l'ITEM, avec l'université de Lausanne, avec les équipes de littérature de la Sorbonne et avec le projet ARTFL de Chicago.

14h00 **Elena Pierazzo** *(Université de Grenoble Alpes)*

▸ **La philologie numérique : une nouvelle méthode ou une nouvelle discipline ?**

L'application des outils numériques au travail du philologue a changé de façon profonde sa méthode, ses objectives et les résultats attendus. Mais est-ce que ces changements seront-ils suffisants pour déclarer qu'il s'agit d'une révolution disciplinaire, ou s'agit-il d'un renouvellement de la méthode qui demeurait presque identique depuis les travaux de Lachmann et de Bédier ?

14h45 **Jean-Baptiste Camps et Lucence Ing** *(École Nationale de Chartes, Université Paris Sciences et Lettres)*

▸ **Collation assistée par ordinateur de témoins de textes en ancien français : défis et perspectives nouvelles**

L'alignement et la collation des variantes des témoins multiples d'une œuvre antique ou médiévale constituent un préalable à l'étude des traditions textuelles et à l'établissement d'un ou plusieurs textes critiques. Lorsqu'elle est réalisée de manière traditionnelle, cette étape est longue et ardue pour le chercheur. Elle conduit en outre à une perte importante d'information, puisque, au cours du processus de collation, l'éditeur choisit généralement de ne retenir qu'une partie des variantes, excluant souvent les variantes graphiques.
Des outils permettant l'automatisation partielle de ce travail de collation, partant de transcriptions exhaustives de tous les témoins, existent. Dans ce cadre, les textes vernaculaires présentent des difficultés particulières, dues notamment à l'importante variation graphique ou à des différences de contenu. Plusieurs pistes seront envisagées dans cette communication pour parer à ces difficultés, comme la présentation de nouvelles formes de visualisation ou la jonction de l'utilisation d'algorithmes comme ceux de CollateX à diverses phases de préparation des données (structuration, lemmatisation…). Un prototype de chaîne de traitement sera esquissé, allant de l'acquisition des textes des témoins à l'annotation des lieux variants, en passant par leur structuration dans un format de référence.

15h45 **Andrés Enrique Arias** *(Universitat de les Illes Balears)*

▸ **New resources for the study of medieval and early modern Spanish biblical translations : the Biblias Hispánicas corpus**

Biblias Hispánicas is a free-access computer tool on the web that features all the Spanish biblical translations composed during the Middle Ages and the sixteenth century, with normalized spelling, lemmatization and POS tagging. This presentation consists of an overview of the main technical and philological issues that arised in the preparation of this corpus and it includes examples of some of its applications for the study of Spanish language and culture in the medieval and Renaissance periods.

16h30 **Marco Presotto** *(Università di Bologna)*

▸ **La edición crítica del teatro clásico español: ¿un modelo exportable?**

El objetivo de la comunicación es el de presentar el contexto actual de ediciones digitales de teatro clásico español a partir de los importantes avances que ha conocido este ámbito de estudios en los últimos decenios. Se tratarán casos específicos, como el de la edición crítica y archivo digital de *La dama boba*, y se intentará establecer hasta qué punto tales proyectos permiten vislumbrar la realización de modelos exportables para la construcción de un corpus digital significativo de este imponente patrimonio de la cultura europea.

17h15 **Isabel Desmet** *(Université Paris 8)*

▸ **Corpus alignés et possibilités de traitement automatique de données pour la description lexicale dans les langues romanes : quelques considérations**

La présente communication se propose de mettre en lumière quelques mécanismes de création lexicale en analysant comment se manifeste ce phénomène dans des écrits spécialisés et de semi-vulgarisation dans les sciences sociales, économiques et financières, dans une perspective comparative entre deux langues romanes, le français et le portugais contemporains.
Par le biais de l'alignement automatique de corpus parallèles, notre approche nous permettra de mettre en lumière différentes stratégies de substitution de termes scientifiques et techniques et donc de synonymie terminologique et d'observer dans quelle mesure ces équivalents surgissent dans des cadres contextuels, textuels et discursifs bien définis.

18h00 **Conclusions et clôture de la journée**