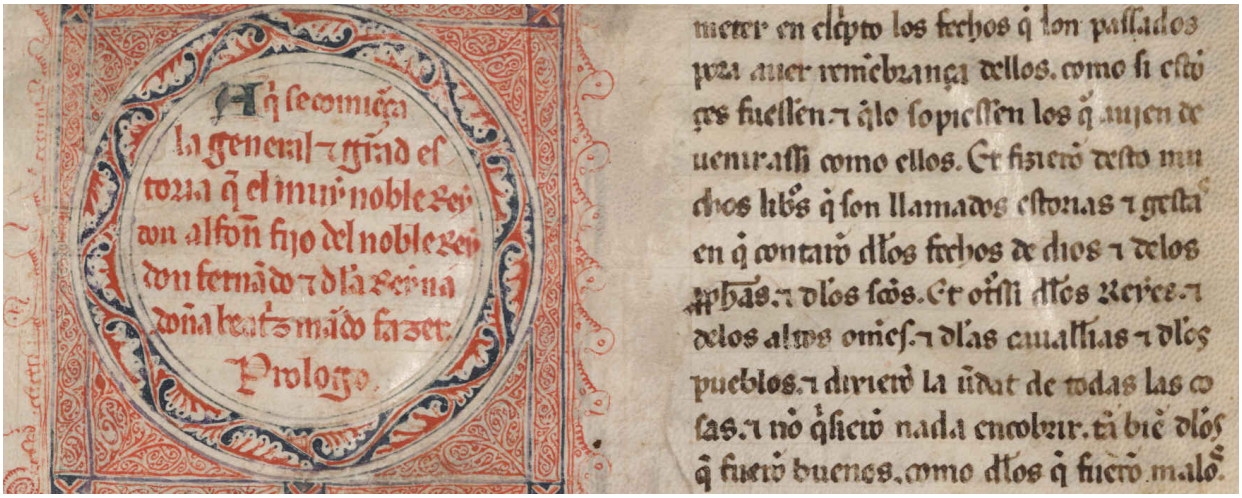

Formation-séminaire sur la transcription automatisée de sources manuscrites avec *eScriptorium*

Organisation : María Díez Yáñez (UCM), Matthias Gille Levenson (EHEHI & ENS de Lyon), Irene Salvo García (UAM)

Casa de Velázquez (Madrid),
23-24 septembre 2021

Présentation

Les jeudi 23 et vendredi 24 septembre 2021 se déroulera à la Casa de Velázquez une formation à la transcription automatisée de l'écriture manuscrite. Rassemblant des intervenants et intervenantes de plusieurs institutions espagnoles et françaises, elle a pour objet la transcription automatisée de sources manuscrites, en anglais HTR (*Handwritten Text Recognition*), via l'outil *eScriptorium* [1], une alternative libre à *Transkribus* [2], développée à l'université PSL en partenariat avec INRIA, qui s'appuie sur l'outil de transcription automatisée *Kraken* [3] et qui montre ses fruits en termes d'efficacité et de disponibilité du code source.

Cette formation sera centrée sur une écriture répandue en Péninsule Ibérique entre le XIII^{è} et le XV^{è} siècle, représentée par un manuscrit du *scriptorium* du roi Alphonse X, le Sage (1221-1284), autour duquel tournera la formation, et à partir duquel sera produit un modèle de reconnaissance automatique qui sera publié par la suite. Il s'agira ici de faire dialoguer philologie et humanités numériques.

1. https ://escripta.hypotheses.org/
2. https ://readcoop.eu/transkribus/
3. https ://dev.clariah.nl/files/dh2019/boa/0673.html

Jeudi 23 septembre

Nous commencerons par une introduction aux écritures de la Péninsule au Moyen Âge, à la tradition paléographique castillane et à ses limites, pour présenter l'écriture du manuscrit choisi, ainsi que son texte et son histoire. Cette première séance sera animée par Leonor Zozaya-Montes (Universidad de Las Palmas de Gran Canarias-CHSC, IATEXT, Universidade de Coimbra). Suivra une intervention d'Irene Salvo García (UAM) pour présenter le texte et l'histoire du manuscrit étudié, et le situer dans la production alphonsine.

Dans un second temps, Peter Stokes (EPHE) et Benjamin Kiessling (PSL), membres de l'équipe de *eScriptorium*, disposeront de l'après-midi pour présenter l'outil, en commençant par une introduction à ce qu'est l'apprentissage supervisé [4] et à ses méthodes. Le logiciel *eScriptorium* sera ensuite présenté, et les formé·es auront deux à trois heures pour transcrire une soixantaine de folios qui aura été répartie entre tous et toutes. Le modèle sera entraîné entre le premier et le deuxième jour.

Vendredi 24 septembre

Pour la première session du second jour, un moment sera consacré à l'évaluation quantitative et qualitative du modèle, pour en déterminer les forces et les faiblesses, et comprendre comment fonctionne un algorithme d'apprentissage supervisé (en se centrant sur les biais éventuels du corpus d'entraînement, *i.e.* le corpus produit par les participants et participantes).

La séance suivante, d'une à deux heures, sera dédiée à la post-acquisition du texte. En effet, en ce qui concerne le castillan médiéval – mais cela peut se généraliser à toutes les langues romanes médiévales –, restent deux problèmes principaux : la segmentation (gestion des « mots » et des espaces dans la phrase : les usages médiévaux sont différents des usages actuels), et la gestion des abréviations. Pour ce qui est des sources volumineuses, la segmentation et la gestion des abréviations doivent idéalement aussi être automatisées sous peine de voir le temps gagné par la transcription automatique perdu par ces deux tâches indispensables et très longues à réaliser à la main. Leonor Zozaya-Montes interviendra sur les méthodes et normes de transcription actuelles, leurs présupposés et leurs limites, puis nous verrons quelles sont les méthodes informatisées pour ces deux tâches, avec les outils les plus récents, et en étudiant deux méthodes possibles : la méthode algorithmique classique et la méthode par apprentissage, chacune ayant ses avantages et ses inconvénients. Les outils de segmentation et de gestion des abréviations pour le castillan médiéval comme pour d'autres langues romanes sont en cours de développement, c'est la raison pour laquelle l'événement que nous organisons tient à la fois de la formation et du séminaire.

Enfin, dans une conférence de clôture, Belén Almeida Cabrejas (Universidad de Alcalá) sera invitée à présenter le projet d'édition CHARTA [5] et le corpus CODEA (*Corpus de Documentos Españoles Anteriores a 1800* [6]), deux projets phares de l'édition et le traitement informatique des textes anciens et de donnés linguistiques dans l'histoire de l'espagnol.

La publication du modèle, dont l'autorité sera partagée entre les participants, les participantes et les organisateurs et organisatrices, donnera lieu à une soumission de *data paper*, éventuellement accompagné d'un retour pédagogique dans une revue spécialisée.

4. https ://fr.wikipedia.org/wiki/Apprentissage_supervisé
5. https ://www.corpuscharta.es/
6. http ://corpuscodea.es/

Organisation et encadrement

Espagne : María Díez Yáñez (UCM), Matthias Gille Levenson (EHEHI/ENS de Lyon), Irene Salvo García (UAM). France : Benjamin Kiessling (PSL), Peter Stokes (EPHE). Deux professeures espagnoles seront invitées à ouvrir et clôturer la formation, Belén Almeida (Universidad de Alcalá) et Leonor Zozaya-Montes (IATEXT, ULPGC - CHSC, Universidade de Coimbra).

Langue de la formation

La formation sera principalement proposée en anglais ; l'introduction, les conférences sur l'histoire du texte étudié et sur les normes de transcription et celle de clôture auront lieu en espagnol. L'intégralité des supports de cours (diapositives, etc.) sera en anglais.

Déroulé de la formation

Jeudi 23 septembre

— **9h00-9h15** : Accueil des participant·es.
— **9h15-12h15** : Introduction - Paléographie - Histoire de l'écriture étudiée (avec pause) (Leonor Zozaya-Montes).
— **12h15-13h00** : Histoire du texte et du manuscrit étudié (Irene Salvo García).
— **13h00-14h15** : Pause déjeuner.
— **14h15-15h00** : L'apprentissage supervisé – introduction, fonctionnement et enjeux scientifiques (Peter Stokes et Benjamin Kiessling).
— **15h00-16h30** : *eScriptorium*.
— **16h30-16h45** : Pause.
— **16h45-19h00** : Atelier pratique : transcription collaborative du manuscrit.
— **19h00** : Fin de la première journée.

Vendredi 24 septembre

— **9h00-9h15** : Accueil des participant·es.
— **9h15-10h15** : Étude du modèle produit : qualités, défauts, biais de corpus éventuels (Peter Stokes et Benjamin Kiessling).
— **10h15-11h45** : Normes de transcriptions et enjeux scientifiques (Leonor Zozaya-Montes).
— **11h45-12h00** : Pause.
— **12h00-13h30** : Après la transcription : segmentation et gestion des abréviations. État de la recherche (Matthias Gille Levenson).
— **13h30-14h45** : Pause déjeuner.
— **14h45-15h15** : Promouvoir utiliser *eScriptorium* dans son université : aspects techniques et financiers (Peter Stokes et Benjamin Kiessling).
— **15h15-17h15** : Conférence de clôture : « *La red CHARTA y el corpus CODEA* » (Belén Almeida)
— **17h15** : Fin de la formation, retours des participant·es.

Candidatures et modalités pratiques

Les candidatures devront être soumises via un formulaire sur le site de la Casa de Velázquez (ici) **avant le 2 août 2021**. Elles devront comprendre :

— un CV synthétique (une page) ;
— une page de présentation des recherches en cours ou à venir du ou de la candidate, comprenant un paragraphe expliquant en quoi la formation peut s'inscrire dans ce projet de recherche.

Nous offrons 20 places environ. La sélection sera rendue publique autour du **9 août 2021**. Une modalité distancielle est envisagée, mais nous donnerons priorité au public sur place, si la situation sanitaire le permet. La session sur l'outil eScriptorium sera ouverte au plus grand nombre, en visioconférence. Un certificat de présence sera distribué à chaque participant·e à la fin de la formation.

Des chambres seront disponibles (aux frais des participants et participantes) à la Casa de Velázquez, en fonction de l'affluence fin septembre. Merci de nous contacter en amont en cas de nécessité.

Contacts

mariadiezy [at] ucm [point] es
matthias.gille-levenson [at] casadevelazquez [point] org
irene.salvo [at] uam [point] es
Lien vers l'événement.

Institutions partenaires et soutien financier

Cette formation est financée par la Casa de Velázquez, la Universidad Complutense de Madrid, la Universidad Autónoma de Madrid et la Communauté Autonome de Madrid (proyecto Canon Hispánico, 2019-T1_HUM-15228).

ÉCOLES FRANÇAISES À L'ÉTRANGER
CASA DE VELÁZQUEZ
UNIVERSIDAD AUTONOMA DE MADRID
Comunidad de Madrid
UNIVERSIDAD COMPLUTENSE MADRID