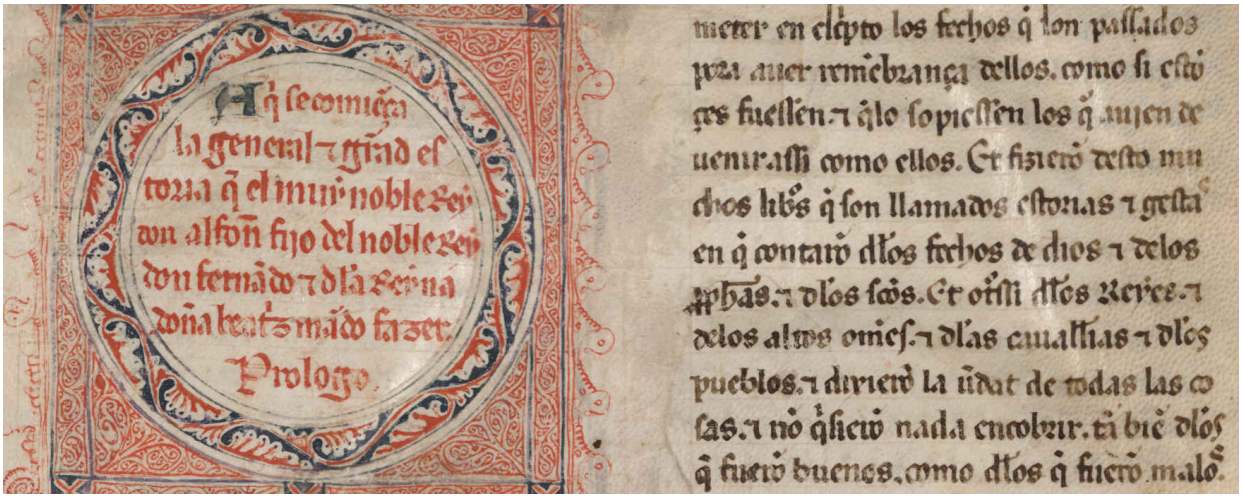

Seminario de formación: transcripción automatizada de fuentes manuscritas con *eScriptorium*

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

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

Presentación

El jueves 23 y viernes 24 de septiembre de 2021 tendrá lugar en la Casa de Velázquez (Madrid) un curso y seminario de formación sobre la transcripción automatizada de fuentes manuscritas. La formación, coordinada por instituciones españolas y francesas, tiene como objetivo la transcripción automatizada de manuscritos, en inglés HTR (*Handwritten Text Recognition*), gracias al programa *eScriptorium*[1]. Este programa ha sido creado en la universidad PSL (Paris Sciences & Lettres) en colaboración con el INRIA (Institut national de recherche en sciences et technologies du numérique). *eScriptorium* es un *software* libre, a diferencia de otros programas como *Transkribus*[2], y está basado en la herramienta de transcripción automatizada *Kraken*[3]. *eScriptorium* ha mostrado resultados sólidos tanto en términos de eficacia y como en la disponibilidad del código fuente.

La formación se centrará en las grafías manuscritas conservadas en la península ibérica entre los siglos XIII y XV, tomando como objeto de estudio un manuscrito del *scriptorium* del rey Alfonso X el Sabio (1221-1284). El manuscrito seleccionado se utilizará para producir un modelo de reconocimiento automático, que será posteriormente publicado. Se pretende, por tanto, establecer un diálogo entre la filología y las humanidades digitales.

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

Jueves 23 de septiembre

El curso comenzará con una introducción a la paleografía peninsular de la Edad Media, con especial atención a los límites del estudio de esta tradición, y una presentación de la escritura del manuscrito elegido. Será impartida por Leonor Zozaya-Montes (Universidad de Las Palmas de Gran Canarias-CHSC, IATEXT, Universidade de Coimbra). Irene Salvo García (UAM) presentará a continuación el texto y la historia del manuscrito alfonsí. En la sesión de la tarde, Benjamin Kiessling (PSL) y Peter Stokes (EPHE), parte del equipo creador de *eScriptorium*, presentarán el programa, comenzando por una introducción a lo que llamamos aprendizaje supervisado[4], a su funcionamiento y a su metodología. Los participantes dispondrán, una vez presentado el programa, de dos o tres horas para transcribir sesenta folios del manuscrito estudiado, que se repartirán previamente entre los asistentes. El modelo de lectura será creado por *eScriptorium* durante la noche del jueves al viernes.

Viernes 24 de septiembre

La jornada comenzará con una evaluación cuantitativa y cualitativa del modelo conseguido el día anterior, con el objetivo de determinar las fortalezas y debilidades de la herramienta y comprender mejor cómo funciona un algoritmo de aprendizaje supervisado (teniendo en cuenta tanto las limitaciones posibles del corpus de entrenamiento como de la transcripción producida por los participantes en la sesión de trabajo del día anterior).

La siguiente sesión, de aproximadamente dos horas, se dedicará a la «post-transcripción» del texto, es decir, al tratamiento de dos de los problemas recurrentes en la transcripción de las lenguas romances medievales: la segmentación de palabras y de espacios en la frase (ya que los usos medievales difieren de los actuales), y el desarrollo y la gestión de las abreviaturas. La consideración de estos dos aspectos es clave para que la lectura automatizada de manuscritos sea productiva y no deba verificarse manualmente una vez extraído el texto. En línea con la primera sesión, Leonor Zozaya-Montes presentará a continuación los métodos y normas de transcripción actuales, sus prejuicios y limitaciones. Matthias Gille Levenson completará la aproximación a la transcripción presentando los métodos informáticos que pueden aplicarse al manuscrito para resolver las cuestiones de segmentación y de abreviación. Para ello recurrirá a las herramientas disponibles más actuales, teniendo en cuenta dos metodologías posibles: el método algorítmico clásico y el método por aprendizaje, con sus respectivas ventajas y desventajas. Las herramientas de segmentación y gestión de abreviaturas para el castellano medieval, como para otras lenguas romances, están en proceso de desarrollo, precisamente por ello la formación propuesta aúna la modalidad de curso y de seminario.

La jornada del viernes terminará con la conferencia de clausura impartida por Belén Almeida Cabrejas (Universidad de Alcalá) que presentará el proyecto de edición CHARTA[5] y el corpus CODEA (*Corpus de Documentos Españoles Anteriores a 1800*) [6]. CHARTA y CODEA son dos proyectos fundamentales en el ámbito de la edición y del tratamiento informático de textos antiguos, así como para la recopilación de datos lingüísticos en la historia del español.

El modelo conseguido durante la formación, cuya autoría será compartida por los participantes y los responsables de la misma, podrá ser objeto de un «*data paper*» que se enviará a una revista especializada en el ámbito del tratamiento digital de textos.

[4] https://es.wikipedia.org/wiki/Aprendizaje_supervisado
[5] https://www.corpuscharta.es/
[6] http://corpuscodea.es/

Organizadores y formadores

España: María Díez Yáñez (UCM), Matthias Gille Levenson (EHEHI/ENS de Lyon), Irene Salvo García (UAM). Francia: Benjamin Kiessling (PSL), Peter Stokes (EPHE). Inaugurarán y clausurarán la formación Belén Almeida (Universidad de Alcalá) y Leonor Zozaya-Montes (IATEXT, ULPGC - CHSC, Universidade de Coimbra).

Idioma de la formación

El idioma de la formación será principalmente el inglés. La introducción, las conferencias sobre la historia del manuscrito seleccionado, las normas de transcripción y la conferencia de clausura se impartirán en español. La lengua de la totalidad de los materiales del curso (diapositivas, material de apoyo, etc...) será el inglés.

Programa del curso

Jueves 23 de septiembre

- 9.00-9.15h: Recepción de los participantes.
- 9.15-12.15h: Introducción - Paleografía – Historia de la escritura del manuscrito seleccionado (Leonor Zozaya-Montes).
- 12.15-13.00h: Historia del texto del manuscrito seleccionado (Irene Salvo García).
- 13.00-14.15h: Comida
- 14.15-15.00h: Introducción, funcionamiento y desafíos científicos (Benjamin Kiessling et Peter Stokes).
- 15.00-16.30h: *eScriptorium*.
- 16.30-16.45: Pausa.
- 16.45-19.00h: Taller práctico: transcripción colectiva del manuscrito.
- 19.00h: Fin de la primera jornada.

Viernes 24 de septiembre

- 9.00-9.15h: Recepción de los participantes.
- 9 .15-10.15h: Estudio del modelo producido: calidad, defectos, sesgos posibles por la parcialidad del corpus (Benjamin Kiessling et Peter Stokes).
- 10.15-11.45h: Normas de transcripción y desafíos científicos (Leonor Zozaya-Montes).
- 11.45-12.00h: Pausa.
- 12.00-13.30h: Después de la transcripción: segmentación y gestión de abreviaturas. Estado de la cuestión (Matthias Gille Levenson).

- 13.30-14.45h: Comida.
- 14.45-15.15h: Cómo promocionar el uso de *eScriptorium* en su propia universidad: aspectos técnicos y financieros (Benjamin Kiessling et Peter Stokes).
- 15 .15-17.15h: Conferencia de clausura: La red CHARTA y el corpus CODEA (Belén Almeida)
- 17.15h: Fin de la formación.

Solicitud y cuestiones prácticas

Las solicitudes deberán enviarse a través de la página web de la Casa de Velázquez (aquí) **antes del 2 de agosto de 2021 (incluido)**. Deben contemplar:

- un CV resumido (una página como máximo),
- una presentación del proyecto de investigación donde figuren los motivos que justifican el interés en asistir al curso, así como las razones que explican la utilidad de la presente formación para la línea de investigación del solicitante.

La formación oferta 20 plazas. La lista de admitidos se publicará el **9 de agosto de 2021**. La participación virtual es posible, aunque se dará prioridad a la presencia de los participantes si el contexto sanitario lo permite. Se emitirán certificados de participación tanto presencial como virtual, con un total de 16 horas de formación, al final de las jornadas.

El alojamiento en la Casa de Velázquez es posible (los gastos correrán a cargo de los interesados) en función de la afluencia registrada en septiembre. Los participantes que necesiten disponer del alojamiento deberán contactar previamente con los organizadores.

Contacto

mariadiezy@ucm.es
matthias.gille-levenson@casadevelazquez.org
irene.salvo@uam.es

Instituciones organizadoras

Esta formación está financiada por la Casa de Velázquez, la Universidad Complutense de Madrid, la Universidad Autónoma de Madrid y la Comunidad de Madrid (proyecto Canon Hispánico, 2019-T1_HUM-15228).

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