BREVE MANUAL PARA LA EDICIÓN Y ANOTACIÓN DE DOCUMENTOS EN CAREXIL-FR

Versión 2.0 (5/5/2024)

Marta López Izquierdo

1. Presentación del proyecto:

El proyecto CAREXIL-FR tiene como objetivo la edición electrónica y la anotación lingüística de una colección de cartas, escritas entre 1939 y 1940, por exiliados y exiliadas españoles internados en campos franceses y destinadas a la CAEERF (Comission d'Aide aux Enfants Espagnols Réfugiés en France). Estas cartas, inéditas hasta la fecha, se conservan en el archivo nacional francés (Archives Nationales, Site de Pierrefitte-sur-Seine, France).

CAREXIL-FR es un proyecto de investigación realizado por el Atelier de Romanités Numériques de la Universidad de París 8 (Laboratoire d'Études Romanes, EA 4378) en colaboración con el Archivo Nacional francés (Archives Nationales)[1]. Tiene por objeto la edición crítica digital y el estudio sociolingüístico de cartas escritas por refugiados españoles internados en campos franceses al término de la Guerra Civil española, conservadas en el archivo francés.

Estas cartas, inéditas hasta la fecha, fueron escritas en su mayoría por mujeres, también por hombres, de diversa edad y procedencia, y están fechadas entre febrero de 1939, coincidiendo con la Retirada, cuando más de 500.000 personas dejan el territorio español huyendo del avance de las tropas franquistas, y el verano de 1940.

Las cartas fueron conservadas en los archivos de la Comission d'Aide aux Enfants Espagnols Réfugiés en France, a quien van dirigidas principalmente. La CAEERF era una organización privada, reconocida por las instituciones francesas y españolas en el exilio, que aportó ayuda material, esencialmente por medio del envío de ropas y calzado, a los miles de españoles internados en campos u otros lugares de acogida más o menos improvisados, con frecuencia en condiciones muy precarias.

Se trata de cartas de petición de auxilio, en las que los firmantes toman la pluma impelidos por la necesidad, dejando constancia en muchos casos de su escasa familiarización con la escritura formal que exige el tipo de texto epistolar que redactan. También se ha conservado para muchas de ellas la copia de la respuesta que la organización les envió, dándonos nuevos detalles sobre la situación vivida por las familias refugiadas.

[1] Y entre 2019 y 2022, con el equipo Études Romanes (EA 369) de la Universidad París Nanterre.

Esta colección de cartas ofrece por consiguiente un material excepcional para el estudio de los repertorios lingüísticos disponibles entre las personas refugiadas, representativas de amplios espectros sociales de la España de la época, y de cómo se solapan las variedades diastráticas y diafásicas dentro de textos pertenecientes a una tipología muy homogénea (cartas de petición formal). De manera más general, estas cartas nos informan sobre la historia de la lengua española en un momento crucial para España.

Más allá del interés lingüístico y sociolingüístico, estas cartas nos permiten acercarnos a la historia, aún insuficientemente conocida y reconocida, de una parte del pueblo español y a las difíciles condiciones en que se produjo su exilio a Francia, del que algunos no volverían nunca.

El proyecto CAREXIL permite el acceso a la edición crítica digital de las cartas conservadas y a su anotación histórica y lingüística para satisfacer las necesidades de filólogos, lingüistas e historiadores interesados.

Este proyecto se inserta en el campo de las Humanidades Digitales, dentro del cual sigue las recomendaciones del consorcio TEI para la edición de textos en formato digital.

Su desarrollo se lleva a cabo a través de la plataforma TEITOK (Maarten Janssen, 2014-).

Los resultados del proceso editorial están inmediatamente disponibles (http://carexil.huma-num.fr) a medida que se van validando las distintas etapas descritas a continuación.

2. Etapas de trabajo para la publicación de las cartas en la plataforma TEITOK:

Para completar el proceso de edición digital de las cartas, deben llevarse a cabo las diferentes tareas que se detallan a continuación:

2. 1. Digitalización de los documentos

2. 2. Transcripción de los documentos en XML-TEI con ayuda de un editor XML (Oxygen)

2. 3. Clasificación de los ficheros resultantes y constitución de un archivo digital

2. 4. Principales etiquetas XML-TEI utilizadas en la edición de las cartas y presentes en la plantilla CAREXIL

2.5. Subida de los ficheros xml y jpg a la plataforma del proyecto:

Se detalla a continuación en qué consiste cada una de estas etapas

2. 1) Digitalización de los documentos y clasificación digital de los mismos:

Esta operación ha de llevarse a cabo directamente en los Archivos Nacionales, sede de Pierrefitte-sur-Seine. Para la mejor calidad del documento digitalizado, se solicitará asiento en la sala prevista para este fin en los archivos (salle de numérisation) y se vigilará que el documento aparezca completo en la fotografía, bien iluminado y con el encuadre lo más ajustado posible, para evitar futuros retoques que hacen perder mucho tiempo.

Para cada documento fotografiado hay que hacer constar la caja, la carpeta y el número de documento que figura generalmente en su margen superior derecho. El formato de la fotografía ha de ser jpg, y ha de hacerse un fichero por cada cara del documento (si el documento está escrito por ambas caras).

Con el fin de localizarlos de manera rápida y precisa, todos los documentos de una misma carpeta se incluirán dentro de una carpeta en el ordenador. Y todas las carpetas de una misma caja en una misma carpeta. Es decir, se respetará en la clasificación digital la clasificación que presentan los documentos en el archivo. Ejemplo:

2. 2) Transcripción de los documentos en formato xml:

La transcripción de los documentos debe ser fiel al estado del texto que se presenta en ellos: se respetan los saltos de línea así como todas las variantes gráficas, de puntuación y de cualquier otro tipo que aparezcan. Se trata de una transcripción semipaleográfica.

La transcripción se lleva a cabo directamente a partir de la plantilla CAREXIL (v. documento CAREXIL_PLANTILLA.xml), que puede editarse en un editor xml como oxygen o en cualquier otro editor de texto xml similar. En caso de no disponer de acceso al editor xml, se puede usar la propia plataforma TEITOK o incluso la plantilla en formato word (CAREXIL_PLANTILLA.docx), respetando el sistema de etiquetado de la plantilla e introduciendo manualmente aquellas etiquetas que sean necesarias. Una vez terminada la transcripción, se realizará su transformación a fichero xml cuando se tenga acceso al programa.

El documento transcrito se grabará siguiendo el sistema de titulación que se describe en el apartado 2. 3.

2. 3) Título de los ficheros:

Los ficheros de texto tendrán el formato xml y las imágenes el formato jpg.

En ambos casos, el título comporta el nombre del proyecto en mayúsculas (CAREXIL), seguido del número de caja (1 a 9), de carpeta y de documento. Cuando el documento está escrito por ambas caras, se precisa el recto (r) o el vuelto (v).

Ejemplos:

CAREXIL_5023013r.xml

CAREXIL_5023013r.jpg

Documento de texto, caja 5, carpeta 23, documento 13 recto.

Documento de imagen, idem.

2. 4) Etiquetado de base presente en la plantilla CAREXIL:

La plantilla CAREXIL contiene las principales etiquetas que permiten anotar los metadatos para cada documento y las principales propiedades de la estructura textual. Estas etiquetas están explicadas en el documento CAREXIL_personalización_TEI.docx, y se recogen a modo de síntesis en el documento CAREXIL_índice_etiquetas.docx.

Presentamos aquí algunas nociones de base del funcionamiento del lenguaje XML-TEI:

1. Las etiquetas siguen el estándar TEI (Text Encoding Initiative), que ofrece un conjunto de etiquetas compartidas por la comunidad investigadora para gran número de contenidos. Puede consultarse el conjunto de etiquetas disponibles y una explicación de su utilización en: https://tei-c.org/ y en una versión para principiantes en: http://teibyexample.org/. El documento ya mencionado (CAREXIL_personalización_TEI.docx) indica qué etiquetas TEI se usan en nuestro proyecto. Este documento puede evolucionar en función de las necesidades que se vayan encontrando. IMPORTANTE: No se pueden cambiar ni añadir las etiquetas individualmente, sino siempre tras previo acuerdo con la coordinadora del proyecto.

2. Las etiquetas xml se sitúan entre paréntesis angulares, y combinan una etiqueta de apertura: <...>, y otra de cierre: </...>.

3. Las etiquetas aportan una información relevante para el tratamiento del texto: sobre su contenido, su forma, su presentación, su clasificación...

4. Las etiquetas pueden ser de dos tipos:

   a. etiquetas dobles: una etiqueta de apertura y una de cierre enmarcando una sección de texto. Ejemplo:

      <salute> Queridas señoras </salute>

   b. etiquetas simples: una sola etiqueta abre y cierra sin enmarcar el texto. Se utilizan para indicar un inicio de línea o de página, por ejemplo.

      <lb/> 'line beginning' (al principio de cada nueva línea)
      <pb/> 'page beginning' (nueva página)

5. Las etiquetas contienen jerarquías: en el primer nivel se sitúan las categorías, que se pueden completar con atributos, en el segundo nivel. Ejemplo:

<lb/><seg type="peroration">Yo espero que Vd haga todo lo que pueda
<lb/>en favor de mi petición tan humanitaria</seg>

donde <seg> = segment, cualquier secuencia de texto, es una categoría
y type="..." es un atributo (@) que permite clasificar el tipo de secuencia textual que se está marcando.
El símbolo @ colocado delante de una etiqueta indica que es un atributo que debe asociarse a una categoría.
Puede notarse que un editor de texto xml asocia automáticamente distintos colores a las categorías (azules) a los atributos (naranjas) y a los contenidos de los atributos (rojo). Pero se puede etiquetar todo en negro sin que esto impida la lectura correcta por el ordenador.

6. Las etiquetas de base de la plantilla CAREXIL anotan los metadatos y la estructura textual. Están pendientes de desarrollo las

etiquetas para anotar la estructura pragmática (borrador disponible en CAREXIL_estructura_pragmática.xml y CAREXIL_estructura_pragmática.docx) y las etiquetas para anotar la estructura lingüística, en particular las variantes sociolingüísticas.

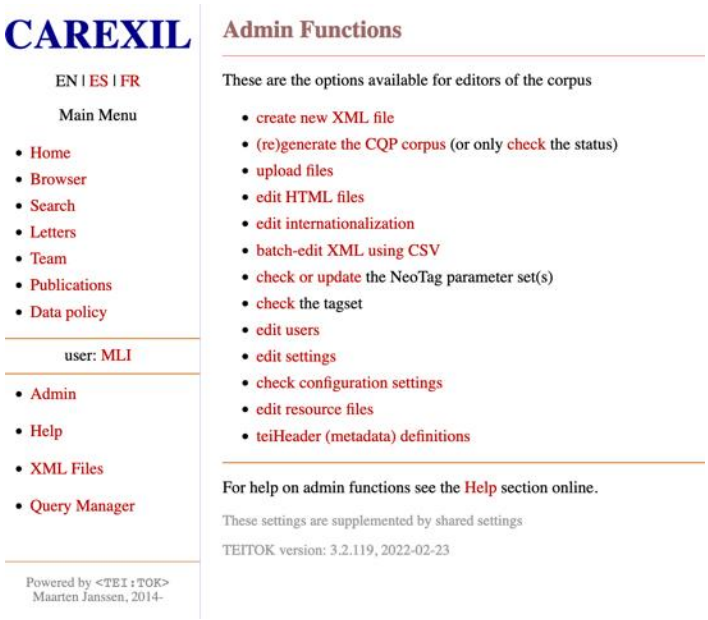
7. En los metadatos, se han marcado entre llaves {…} los contenidios que tienen que actualizar los colaboradores (los demás contenidos son iguales para todas las cartas y ya están incluidos en la plantilla). Esto permite reconocer más rápidamente los campos que deben rellenarse en esta sección del xml (con una búsqueda automática con la función Find de Oxygen, por ejemplo).

2. 5) Subida a la plataforma TEITOK :

Cada colaborador autorizado dispondrá de un identificador para conectarse a la plataforma y poder realizar la tarea de editar un texto. El proceso debe seguir varias etapas, en este orden:

1. cargar el documento xml en la plataforma (comando: admin > upload files > TEI XML files).

<img scr: http://carexil.huma-num.fr/Facsimile/image004.jpg



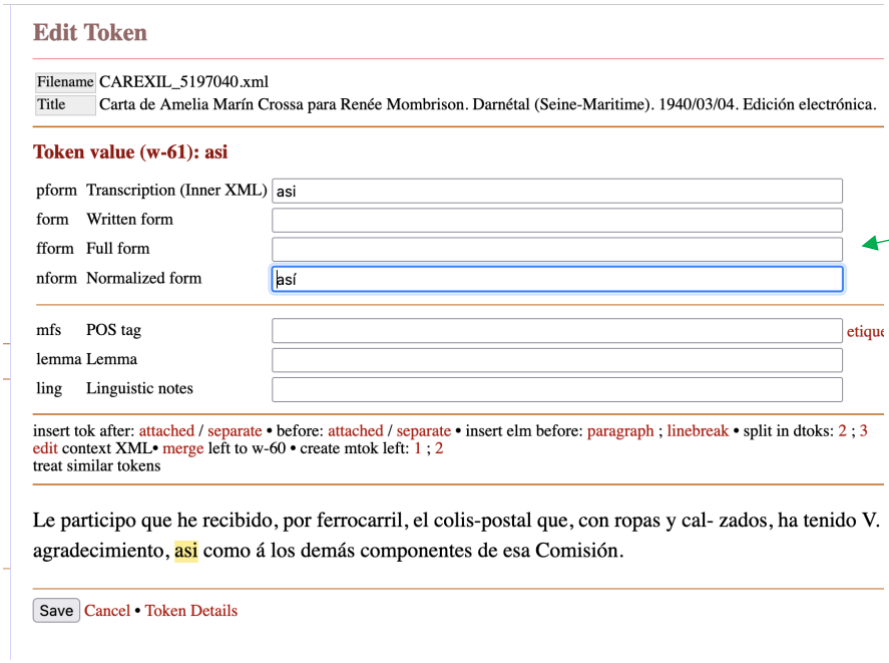
2. cargar la imagen jpg del facsímil: deslizar el ratón al final de la página y pinchar en: “select a different type file”, o bien volver a la función anterior : Upload files > Facsimile Images.

3. en el menú inicial, ir a XML Files, y buscar por orden de número el fichero que se acaba de subir: comprobar el estado del documento y que la transcripción no contiene errores.

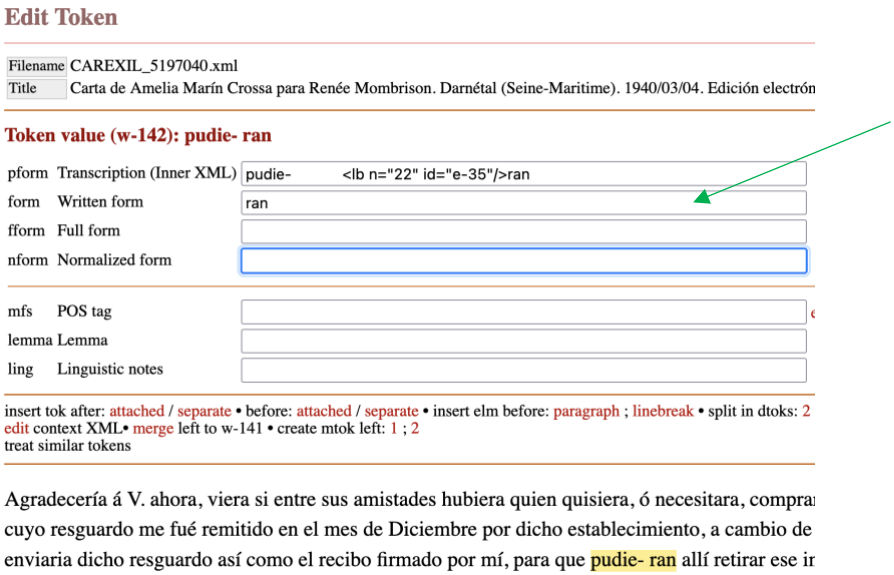
4. revisión 1: otro colaborador del proyecto revisa la transcripción. IMPORTANTE: Solo cuando se ha dado el visto bueno puede pasarse a la etapa 5.

5. tokenización: operación automática que se lleva a cabo pinchando en el comando “tokenize the text” de la plataforma (en la zona inferior de la carta que se está editando, aparece el menú: Admin options). Consiste en atribuir a cada unidad una etiqueta de identificación de token (<tok>) que permitirá posteriormente su tratamiento automático.

6. normalización: en el menú anterior, se pincha ahora en “normalize text” y se comprueba que la normalización se ha efectuado correctamente. Cuando no es el caso, se seleccionan las palabras que no presentan una forma estándar y se pincha dos veces, hasta que se abre un menú de diálogo donde se puede corregir el etiquetado. En la línea "normalized form" se añade la forma estándar conforme al español europeo contemporáneo.



7. para las palabras separadas por guión (o sin guión) al final de la línea (pudie- ran) es necesario realizar una operación de fusión (‘merge’) para que se reconozcan como una sola unidad: pinchar sobre el segmento final de la palabra y en el cuadro de diálogo anterior, seleccionar “merge left”. Aceptar y repetir la operación hasta que aparezca la palabra completa en el cuadro de diálogo. En “written form”, escribir la forma completa sin guión: pudieran.



8. revisión 2: un nuevo revisor verifica la normalización y las operaciones especiales antes de pasar a la última etapa.

9. etiquetado POS ("part-of-speech") y lematización: pinchar en el comando indicado al final de la página Admin opt: "(Pre)tag this text with POS (and lemma)".

10. comprobación de las etiquetas y corrección cuando sea necesario: pinchar la palabra en la que hay un error de etiquetado y corregir la línea POS o lema, según el caso.

11. revisión 3: revisión final.