DOI 10.14672/2.2025.3218

FLORENCIA FERRANTE
CHIARA VALENTE
ANA LOURDES DE HÉRIZ

EL LEMARIO ITALIANO-ESPAÑOL DEL *DICCIONARIO DE FALTRIQUERA* DE CORMON Y MANNI (1805): TRANSCRIPCIÓN AUTOMÁTICA Y CLAVES ECDÓTICAS

Università degli Studi di Genova

Resumen

El *Diccionario de faltriquera* italiano-español y español-italiano de Cormon y Manni (1805) es una de las obras transcritas para su digitalización e inclusión en el *Tesoro TELEI*. El artículo expone las características tipográficas de la macroestructura y microestructura del lemario italiano-español que han supuesto una dificultad en la transcripción automática, así como las soluciones que se han tomado tras el cotejo de fuentes primarias y obras publicadas posteriormente en las que este diccionario influyó.

palabras clave: lexicografía bilingüe, italiano-español, transcripción automática, Cormon y Manni

Abstract

The Italian-Spanish section of Cormon and Manni's Dictionary *(1805): automatic transcription and ecdotic clues*

The Diccionario de faltriquera *italiano-español y español-italiano by Cormon and Manni (1805) is one of the dictionaries transcribed for digitalisation and publication in the* Tesoro TELEI. *This paper describes the typographical characteristics of the macrostructure and microstructure of the Italian-Spanish section, which have represented a difficulty in the automatic transcription, as well as the solutions that have been taken, also after a comparison with the primary sources and later published dictionaries on which this dictionary had an influence.*

keywords: bilingual lexicography, Italian-Spanish, automatic transcription, Cormon & Manni

Recibido: 28 de marzo de 2025
Aceptación definitiva: 25 de junio de 2025

1. Introducción

Las páginas que siguen[1] presentan una investigación aplicada que se inscribe en el marco de un proyecto[2] cuyo objetivo general es la retrodigitalización del patrimonio lexicográfico bilingüe italiano-español (s. XVI-XX) para su encontrabilidad, reutilización, interoperabilidad y accesibilidad universales, según los principios FAIR (Wilkinson *et al.* 2016). El proyecto se sitúa en una tradición reciente de estudios que intentan combinar, en lo que se ha denominado un verdadero *paradigm shift*, la lexicografía, la ontología y las herramientas de la lingüística computacional en las Humanidades Digitales (Costa *et al.* 2021).

El término 'retrodigitalización' suele hacer referencia, por un lado, al proceso de transformación en documento digital legible por máquina (PDF, por ejemplo) de un objeto textual publicado en papel antes del advenimiento de las Humanidades Digitales; por el otro, 'retrodigitalizar' significa también crear, a partir de la fuente antigua en papel, un objeto digital en formato de archivo estructurado y codificado en XML, que permita la interoperabilidad con el usuario y la búsqueda y selección de datos y metadatos específicos de cada objeto a partir de una interfaz interactiva (Castillo Peña 2020, Nalesso 2024). Huelga decir que el tipo de retrodigitalización al que aspira el proyecto en curso es del segundo tipo, o sea el diseño y la creación de un tesoro de la lexicografía bilingüe italiano-español en el que los objetos textuales digitales permitan a cualquier tipo de usuario la realización de búsquedas parametrizadas complejas de información semántica y estructural, así como también la comparación y alineación de los contenidos seleccionados entre los distintos objetos digitales.

El primer paso de la investigación ha consistido, por tanto, en el desarrollo de un flujo de trabajo para experimentar un protocolo de retrodigitalización de un corpus reducido de obras lexicográficas bilingües, escogidas según cuatro criterios fundamentales descritos y justificados en Nalesso (2024: 139): cronológico, tipológico, historiográfico y metodológico.

Cada una de las universidades que participan en el proyecto TELEI (Padua, Génova, Bolonia, Pisa, Turín y Verona) constituyen sendos grupos de investiga-

1 Las tres autoras han colaborado a la par en el proyecto de este artículo y en la revisión final. No obstante, cabe precisar que las secciones §1 y 2 se deben a Florencia Ferrante, las secciones §3, 4, 5 y 6 a Chiara Valente y las secciones §7 y 8 a Ana Lourdes de Hériz.

2 El proyecto lleva como título *Un nuevo espacio digital para el patrimonio lexicográfico: el 'Tesoro digital de la lexicografía bilingüe español-italiano'*, tesoro digital representado por la sigla TELEI. Esta investigación está siendo financiada por La Unión Europea-Next Generation EU, Missione 4 Componente 1 CUP C53D23004010006 / PRIN 2022 MUR 20229W73WR.

ción que se ocupan del proceso integral de retrodigitalización de una o más de las obras lexicográficas seleccionadas para el *Tesoro* digital. Así, cada grupo tiene a cargo objetos textuales distintos con características específicas. El grupo de la Universidad de Génova, cuyo protocolo de trabajo se presenta en estas páginas, tiene asignada la retrodigitalización de los dos volúmenes del *Diccionario de faltriquera italiano-español y español-italiano* de Cormon y Manni (primera edición en Lyon, 1805)[3].

A la selección de los materiales, la asignación de estos a cada uno de los grupos y el estudio de las principales características filológicas y metalexicográficas del corpus, sigue el proceso de retrodigitalización propiamente dicho. Obviamente, estaba previsto que, antes de la fase de codificación y modelización de los textos para su etiquetación semántica en formato XML según las directrices TEI Lex-0[4], se llevara a cabo la transcripción de los diccionarios, la mayoría de los cuales ya estaban disponibles en formato digital (.pdf o .jpg, por ejemplo), para su transformación en documentos legibles por máquina (en formatos como .txt o .doc, por ejemplo). Para ello, cuando se ha optado por una transcripción automática y no manual, se ha utilizado el programa de reconocimiento y transcripción de documentos antiguos *Transkribus*[5], desarrollado por la Universidad de Innsbruck en el marco del proyecto READ[6] (Rabus 2019) y hoy ampliamente utilizado en los estudios filológico-literarios.

Transkribus, sin embargo, no se ha experimentado hasta ahora exhaustivamente para el reconocimiento y la transcripción de materiales lexicográficos. Estos

3 A este diccionario se le han dedicado extensos estudios por ser uno de los textos fundamentales de la lexicografía italoespañola del siglo XIX (San Vicente 2010; Martínez Egido 2008, 2010; Alvar Ezquerra 2010). Remitimos a estas referencias bibliográficas a quien no conozca esta obra, que atravesó con muchas reimpresiones y ediciones revisadas todo el siglo XIX y que fue fuente primaria de otras obras lexicográficas.

4 TEI Lex-0 es una especificación técnica y también un conjunto de recomendaciones generales para codificar diccionarios legibles por máquina. Está basado en las directrices generales de la *Text Encoding Initiative* (TEI) y se ofrece como una personalización de sus directrices generales. El principal propósito del proyecto TEI Lex-0 es el de proporcionar un esquema de codificación simplificado y estandarizado para cualquier tipo de recurso lexicográfico antiguo o moderno, así como también para otros tipos de texto con características similares (enciclopedias, catálogos), que sea fácil de adoptar y utilizar por parte de los editores y que permita, además, la interoperabilidad y la reutilización de los contenidos.

5 Transkribus© READ-COOP SCE [22/03/2025] <https://www.transkribus.org/>.

6 *European Commission Horizon 2020 Research and Innovation Programme*, *Recognition and Enrichment of Archival Documents* (READ [22/03/2025] <https://readcoop.org/>).

tienen, como es sabido, una serie de características tipográficas, macro y microestructurales específicas, que no siempre son de fácil procesamiento para este tipo de *software*. Asimismo, el diccionario asignado a nuestro grupo de investigación, el *Diccionario de faltriquera italiano-español y español-italiano* de Cormon y Manni (Lyon 1805) ha supuesto, por una serie de motivos de los que daremos cuenta enseguida, un desafío aun mayor para la transcripción automática.

A continuación, tras un breve estado de la cuestión acerca del uso de *Transkribus* en proyectos de corte lexicográfico (§2), se dará cuenta de las dificultades para el reconocimiento y la transcripción del texto macro y microestructural de la obra y de las soluciones adoptadas, para las que ha sido a menudo necesario tomar en consideración la historia editorial y filológica de esta obra.

2. *Transkribus* y diccionarios

La creación, desarrollo y validación del protocolo de retrodigitalización de materiales lexicográficos bilingües de italiano-español es uno de los objetivos específicos del proyecto TELEI. Uno de los pasos esenciales para generar un flujo de trabajo eficiente es la elección de las herramientas informáticas adecuadas. Entre los recursos informáticos más difundidos hoy en día para la retrodigitalización de textos antiguos y modernos se encuentran los siguientes: el ya mencionado *software Transkribus*, especializado en el reconocimiento y transcripción automática de documentos históricos impresos y manuscritos, y GROBID[7], un programa también configurado para el reconocimiento y estructuración automática de distintos tipos de documentos (PDF, por ejemplo) en un archivo en formato XML/TEI. Sobre este programa se ha desarrollado, en los últimos años, un módulo específico para la codificación en formato XML/TEI Lex-0 de cualquier tipo de documento lexicográfico en formato digital, conocido como GROBID *Dictionaries* (Khemakhem *et al.*2017).

La elección de *Transkribus* para el reconocimiento y transcripción del corpus experimental de nuestro proyecto fue una decisión ponderada por todos los grupos de esta investigación, en la medida en que nos permite registrar y evaluar la utilidad, los problemas, las ventajas y desventajas de su aplicación para la digitalización de documentos lexicográficos y sentar un precedente para investigaciones y proyectos futuros. El proceso compartido de la experimentación, las dificultades

7 GROBID (GeneRation Of Bibliographic Data) [22/03/2025] <https://grobid.readthedocs.io/en/latest/>.

CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

encontradas, así como las soluciones adoptadas, son parte, pues, de los resultados de la investigación en curso.

Transkribus, como ya se ha indicado más arriba, no ha sido utilizado masivamente para la transcripción y digitalización de textos lexicográficos. Entre los escasos empleos de este *software* en proyectos de digitalización de diccionarios o textos similares como bibliografías, se puede mencionar el estudio de Lindemann *et al*. (2018), quienes integran el uso de *Transkribus* con GROBID *Dictionaries*. En otro proyecto similar más reciente, Lindemann y Alonso (2021) presentan un flujo de trabajo para la digitalización de diccionarios históricos, pero optan por utilizar *Kraken*[8], un programa con una función equivalente a la de *Transkribus* pero con características y especificaciones distintas, más adecuadas para resolver cierto tipo de requerimientos infraestructurales. Otros dos grandes proyectos de mención obligada de retrodigitalización de obras lexicográficas llevados a cabo en los últimos años son el *Nenúfar Project*[9] y el *MorDigital Project*[10]. El *Nenúfar Project*, cuyo objetivo es la retrodigitalización de distintas ediciones del diccionario monolingüe *Petit Larousse Illustré*, utiliza también el *software* GROBID *Dictionaries* sobre un corpus del que ya se habían obtenido archivos legibles por máquina (Bohbot *et al.* 2018). *El MorDigital Project*, con el fin de retrodigitalizar las primeras tres ediciones del *Diccionario da Lingua portugueza* de Antonio de Morais (publicado en 1789), utiliza también GROBID *Dictionaries* para generar directamente, a partir de los documentos en .pdf, un archivo XML estructurado en formato TEI Lex-0 (Costa *et al.* 2021). Como puede verse, los proyectos más recientes de retrodigitalización a gran escala del patrimonio lexicográfico monolingüe en lenguas románicas parecen haber optado por otras herramientas de transcripción, digitalización y codificación como *Kraken* o GROBID *Dictionaries* en lugar de *Transkribus*.

Las dificultades más frecuentes a las que se suele hacer referencia en los estudios recién citados, por lo que atañe a la transcripción automática para la posterior codificación de textos lexicográficos, son: i) la frecuente baja calidad de las imágenes del documento fuente (Khemakhem *et al.* 2019); ii) las características tipográficas del texto antiguo, dado que suelen ser fuente de problemas para el reconocimiento automático de caracteres; y iii) respecto a algunos casos específicos, la escasa correspondencia entre la macro y microestructura del diccionario y las etiquetas establecidas para la codificación (Lindemann, Alonso 2021).

8 [22/03/2025] <https://kraken.re/main/index.html>.

9 [22/03/2025] <https://nenufar.huma-num.fr/presentation/>.

10 [22/03/2025] <mordigital.fcsh.unl.pt/en/homepage/>.

3. Características tipográficas de valor metalexicográfico

El lemario que ya ha sido transcrito completamente (tomo I, del italiano al español) es una de las dos partes de lo que el título de la 1ª edición de la obra presenta como un diccionario de bolsillo ("de faltriquera"). Este formato respondía a necesidades de tipo comercial y escolar (Cazorla Vivas 2010: 32-33) y obligaba a la maquetación de una macroestructura y una microestructura peculiares. Si se observa una página del lemario, se nota enseguida una disposición de los elementos de las entradas que prima el ahorro de espacio, con el objetivo de incluir el mayor número de palabras posible en una obra de solo 14 cm de largo.

El tomo I contiene un lemario de 422 páginas; a este se añaden los siguientes paratextos iniciales, que no hemos transcrito por ahora: el prólogo (páginas I-II), el apartado dedicado a las "conjugaciones de los verbos regulares e irregulares" (III-XXII), la lista del catálogo "d'alcuni libri italiani e spagnuoli che si trovano nella stessa libreria" y la lista "explicación de las señales y abreviaturas usadas en este tomo"; lista que se ha tomado en consideración cuando han surgido dudas de interpretación del texto producido por el escaneado automático.

Por lo que se refiere al formato de las páginas del lemario, el texto está repartido en tres columnas; encima de cada una de ellas aparecen las primeras tres iniciales-guía de la primera palabra lematizada en dicha columna. En lo alto de la página, a la derecha o a la izquierda (en las pares o impares, respectivamente), están las cifras arábigas de numeración de cada página. En el pie de página, a la izquierda, cada dieciséis páginas aparece la inscripción *Tomo I*. En la primera de cada pliego (16º), empieza una serie de signaturas tipográficas representadas en orden alfabético por una letra y un número arábigo (A, A2, A3, A4, B, B2, etc.).

Cada columna incluye un promedio de unas 30 entradas, es decir, alrededor de 100 entradas por página, en un volumen de dimensiones muy reducidas. Para aprovechar al máximo este espacio limitado, la microestructura presenta algunas estrategias peculiares que, aunque no fueran innovaciones en la lexicografía del tiempo −como se detallará más abajo−, tampoco estaban muy difundidas.

La microestructura básica del artículo lexicográfico más simple se desarrolla del siguiente modo: el lema está escrito en redondilla y solo con la inicial mayúscula; le sigue una coma que lo divide de las abreviaturas de la categoría gramatical, separadas cada una de ellas por un punto[11]. Tras estas sigue, en cursiva, el equivalente de traducción. Si se propone más de un equivalente de traducción en

11 Se ha observado un uso no siempre coherente de las abreviaturas; algunas variantes no aparecen en la lista incluida en el volumen del tomo I del diccionario.

CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

las entradas monosémicas o con más de una acepción de significado, se separan con una coma (es lo que se observa más abajo en la imagen 1 de los artículos de *Faccenda* y *Faccendiere*).

A partir de esta estructura básica, se pueden encontrar diferentes variantes. Si el artículo comprende más de una acepción de significado con sus correspondientes equivalentes de traducción (véase *Laidezza*, imagen 2), se separan mediante el símbolo de la doble daga (‡, explicada en la lista inicial de "señales" del diccionario como "señal para separar las varias acepciones de una misma voz"). Este símbolo se emplea también para diferenciar dentro de un artículo las subentradas por categoría gramatical; en este caso, a la doble daga le sigue la abreviatura de la nueva categoría gramatical (v. imagen 3 de *Preciso*, adjetivo y adverbio en italiano).

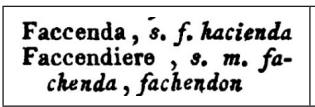
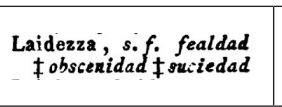
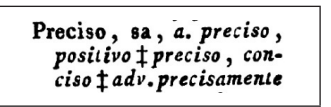
		

Imagen 1	Imagen 2	Imagen 3

Además de este uso del símbolo de la doble daga, resulta peculiar el recurso al signo gráfico de la llave de cierre en forma de arco para recoger entradas de lemas que comparten el mismo equivalente de traducción, pero que tienen una estructura morfológica diferente (normalmente con variación sufijal). A la izquierda de la llave se listan verticalmente los lemas y, a la derecha, se indica la categoría gramatical seguida del equivalente o los equivalentes de traducción (v. imagen 4, lematización de *Importunezza* e *Importunità*). Si cada lema tiene una categoría gramatical diferente (o incluso solo una diferencia de género) se indica antes de la apertura de la llave tras cada forma de las entradas (v. imagen 5 de *Potagione* y *Potamento*).

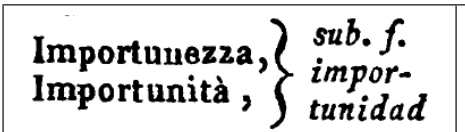
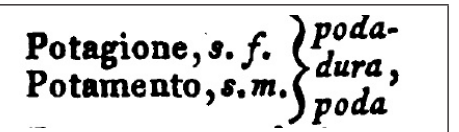
	

Imagen 4	Imagen 5

En todo el texto de las tres columnas se puede apreciar además el uso muy frecuente de guiones que sirven para cortar una palabra y completarla en la línea siguiente. Este recurso tiene el evidente objetivo de aprovechar el espacio de las

líneas hasta el final. Sin embargo, y es esta otra de las peculiaridades tipográficas de la obra, muchas veces las palabras equivalentes de traducción se completan en la línea de arriba, si el espacio sobrante del artículo anterior lo permite, separando esa fracción de la palabra cortada del texto de la línea superior mediante el símbolo de apertura de un corchete. Es una estrategia que permite aprovechar al máximo el espacio, pero que hace más difícil la lectura, como puede verse en la secuencia de 4 lemas *Implicato-Implicitamente* de la imagen 6.

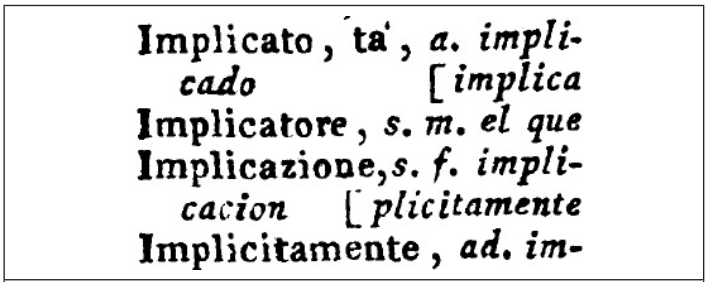

Implicato, ta, a. impli-
cado [implica
Implicatore, s. m. el que
Implicazione, s. f. impli-
cacion [plicitamente
Implicitamente, ad. im-

Imagen 6

Con todo ello no se debe pensar que consideramos el uso de estos símbolos y estrategias como algo extraño en la lexicografía de principios del s. XIX; sí, en cambio, que no era lo más frecuente. Por ejemplo, el uso del corchete de apertura para terminar una entrada en el espacio sobrante de la anterior ya se había visto en otros diccionarios compactos con espacio reducido como el de Gattel *Nuevo diccionario portatil español y francés*, de 1798 (Bruña Cuevas 2008: 54; Cazorla Vivas 2002: 311-28). El arco del símbolo de la llave que une varias formas flexivas de un lema se usaba también en diccionarios no portátiles como el *Dictionaire italien, latin e françois* del Abad Antonini, publicado en París en 1743 (Lillo 2019: 152-54; Mormile 1993: 52-6). En el *Nuevo diccionario portatil español é ingles* de Gattel, publicado en 1803 (Lombardero Caparrós 2015: 149) −dos años antes que el Cormon y Manni de nuestro proyecto− coinciden ambas estrategias de ahorro de espacio, las llaves y los corchetes que encierran la continuación del artículo en la línea de arriba. Hemos dedicado un espacio a exponer estos recursos tipográficos por su valor metalexicográfico y por las complicaciones que han provocado en la fase de transcripción automática del texto, pues, como se verá en §5, *Transkribus* no los ha reconocido automáticamente.

CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

4. El proceso de transcripción del tomo I del diccionario

Tal como se ha anticipado, la transcripción del diccionario seleccionado para el proyecto se ha llevado a cabo con el *software* en línea *Transkribus*. A continuación listaremos sucintamente los pasos que se han seguido para la transcripción, control, revisión y fijación definitiva del texto del tomo I, deteniéndonos luego en algunos detalles que consideramos que pueden ser interesantes.

4.1 *Pasos realizados*

1. Revisión de todos los ejemplares del tomo I a disposición en formato .pdf y elección del que se iba a transcribir.
2. Selección de las páginas del volumen que se iban a cargar en el *software* y subida de archivos PDF a *Transkribus*.
3. Transcripción con *Transkribus* en este orden: segmentación de la página en regiones (correspondientes a las tres columnas del lemario de cada página), escaneo, reconocimiento de las líneas y del texto mediante modelo público y generación de transcripción automática. Este proceso se ha aplicado inicialmente a las primeras 120 páginas del lemario (letras A, B, C, D y E), cuyo texto generado automáticamente se ha ido corrigiendo para educar al sistema de escaneo respecto a algunos errores de reconocimiento de letras y otras peculiaridades (como las presentadas en §5 y §6).
4. Creación de varios modelos privados de HTR (*Handwritten Text Recognition*) y aplicación al lemario segmentado en diferentes archivos PDF.

4.1.1 Elección del ejemplar para la transcripción

Los ejemplares del tomo I de la edición de 1805 de este diccionario que teníamos a disposición en formato .pdf son de las siguientes bibliotecas[12]: Biblioteca de la Universidad Complutense de Madrid (a partir de ahora, ejemplar BUCM),

12 Nos basamos en los sellos que se pueden observar en las portadas u otras páginas de los ejemplares y en el hecho de que los números de registro que aparecen escritos en su interior coinciden con los que constan en los catálogos. Junto con la sigla de las bibliotecas mencionadas, indicamos la signatura del catálogo de estos ejemplares: BUCM (D 50877); BNF (MAGL. 7.8.5 0001), BNE (U/2548). Sobre el ejemplar de la Biblioteca municipal de Forlì, veáse la siguiente nota.

Biblioteca Nazionale di Firenze (ejemplar BNF), Biblioteca Nacional de España (BNE) y Biblioteca Municipale di Forlì (BMF)[13].

Para la transcripción automática decidimos recurrir, tras varias pruebas, al ejemplar de la BUCM. Se tomó esta decisión por razones prácticas: ante todo, se descartaron los ejemplares de la BNE y de la BMF por contener ambos dos páginas del volumen en cada página del PDF (en formato apaisado). El hecho de presentar dos páginas del diccionario en una misma del PDF dificultaba aún más el reconocimiento de las regiones por parte del programa. De los otros dos ejemplares, se ha seleccionado el BUCM porque parecía que *Transkribus* lo escaneaba generando menos errores. Esta elección, sin embargo, no debe llevar a la conclusión de que la transcripción se basa únicamente en este ejemplar; para resolver todas las dudas que surgieron a lo largo de la revisión de la transcripción, hubo que cotejar los diferentes ejemplares en PDF.

4.1.2 Selección de los apartados para cargar en el *software* y subida de estos a *Transkribus*

Tras elegir el ejemplar para transcribir, se dividió en archivos más pequeños, para agilizar el proceso de transcripción y trabajar con calas "letra por letra", o sea archivos que contuvieran todos los lemas que iniciaban por la misma letra. El número de páginas de cada uno de estos archivos no coincide ya que depende de la cantidad de lemas que empiezan por cada letra. El primer archivo era el más extenso (de unas 120 páginas) y contenía la cala que va de la letra inicial A hasta la E (incluida)[14].

13 La biblioteca en la que está guardado el ejemplar que llamamos BMF es la Aurelio Saffi de Forlì, actualmente cerrada por obras. El ejemplar forma parte del fondo antiguo, cuyo catálogo no ha sido digitalizado. El PDF de este tomo I del diccionario está disponible en el portal Contrastiva.it, de San Vicente (dir.).

14 En el momento de esa fase del trabajo se pensó que entrenar al *software* con un archivo muy extenso perfectamente revisado iba a facilitar que no se produjeran tantos errores al escanear los archivos más breves de las siguientes letras. La experiencia que se está llevando a cabo con la transcripción del tomo II nos ha llevado a entender que conviene educar al *software* al principio con archivos breves que se van convirtiendo en modelos cada vez más extensos a medida que se les añade el siguiente archivo revisado y aprobado.

CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

4.1.3 Transcripción con *Transkribus*

Esta fase del trabajo se puede sintetizar con los siguientes pasos realizados en cada una de las páginas de un archivo PDF: segmentación de la página en regiones, reconocimiento de líneas y del texto con un modelo público y correcciones de las imperfecciones y otras peculiaridades que queríamos enseñar al programa. Tras haber subido a la web de *Transkribus* los archivos correspondientes a las letras, se ha iniciado el reconocimiento de texto utilizando el modelo público llamado *Transkribus print M1*. Se ha elegido este modelo porque ya lo habían experimentado otros grupos de investigación del proyecto TELEI y habían llegado a la conclusión de que funciona bien con textos en español e italiano de los siglos XVIII y XIX. Al principio de la transcripción de las primeras páginas, la división en regiones se llevó a cabo de forma manual, utilizando el instrumento *Regions* de *Transkribus*; luego dejó de ser necesario.

4.1.4 Creación de varios modelos privados de HTR y aplicación a los restantes archivos

A continuación, se creó el primer modelo privado de HTR, basado en el modelo público antes mencionado y en la transcripción revisada y aprobada de todos los artículos de las letras A-E. Dicho primer modelo de texto se aplicó a los archivos de los lemas con las letras iniciales F, G, H, I, J y L. El reconocimiento del texto y la transcripción automática con este primer modelo dio algunos resultados interesantes: el sistema reconocía el símbolo que separa en este diccionario las acepciones con la doble daga (‡) y lo transcribía con el símbolo que habíamos decidido adoptar (//)[15]. Asimismo, reconocía mejor que el modelo público la letra <ñ>. Sin embargo, con este primer modelo privado seguían produciéndose errores de reconocimiento de algunas letras −sobre todo mayúsculas− y de varios signos de puntuación.

15 Una breve nota para justificar esta equivalencia (//) en la transcripción de la doble daga. Este símbolo tipográfico, de antigua tradición y con múltiples valores según fuera el impresor que lo usaba, puede ser representado actualmente con programas de escritura de texto mediante el código U+2021, o sea, así: ‡. En el momento en que se tomaron algunas decisiones respecto a la revisión y aprobación del texto generado para poder ir creando modelos, se pensó que este símbolo podía ser difícilmente aceptado por el sistema de escaneo automático y por ello se adoptó la doble barra inclinada. Nada impide intervenir luego en el archivo global extraído de *Transkribus* en un archivo TXT con una operación busca/sustituye y cambiar todas las dobles barras inclinadas por una doble daga ‡.

A partir de ahí se creó un segundo modelo que se aplicó a los archivos de las letras iniciales M y N. Cabe decir que, en general, dicho modelo funcionaba mejor en el reconocimiento de las letras y, sobre todo, de las comas y los guiones, que antes se confundían o no se transcribían. El tercer modelo, creado a partir de la transcripción revisada de los lemas de las letras iniciales M y N, no funcionaba bien, probablemente por algún error humano cometido en la fase de creación del modelo. Por ello, hubo que preparar un cuarto modelo con una técnica diferente para mejorar el reconocimiento de las letras mayúsculas iniciales: para generar este modelo no nos hemos basado en archivos enteros del lemario con todos los lemas con varias letras iniciales como hasta ese momento, sino que tomamos como modelo las primeras tres páginas transcritas y revisadas de las letras iniciales O, P y Q, para que el sistema tuviera una muestra de las letras mayúsculas que había que reconocer y pudiera aplicar el nuevo conocimiento a las páginas por transcribir con estas mismas letras mayúsculas iniciales. Con este método se han obtenido buenos resultados. El nuevo modelo reconoce bien las letras mayúsculas, ha aprendido a no leer la parte de la línea que sigue a los corchetes, ya que en todas las correcciones manuales esta parte se había borrado y añadido manualmente en la línea de abajo; obviamente, esta última operación tuvo que seguir siendo manual durante todo el proceso.

El quinto modelo se ha creado con la misma técnica, con las primeras páginas de las letras R, S y T, y se ha aplicado luego al resto de las páginas de dichas letras. Lo mismo se ha hecho con el sexto y último modelo y las letras U, V, X y Z. Estos últimos dos modelos se han creado con el objetivo de mejorar cada vez más el reconocimiento del texto. Además, se ha observado un evidente progreso en la lectura de las palabras, aun cuando algunas letras eran casi ilegibles o estaban borradas, como si el sistema, conociendo la lengua de esa parte del texto, pudiera reconstruir la palabra a pesar de estar incompleta.

Por lo que se refiere al entrenamiento de *Transkribus* con modelos, cabe añadir que también creamos un modelo *layout*, con el objetivo de que leyera las tres regiones de las tres columnas de manera automática, sin tener que enmarcarlas manualmente antes de empezar el escaneo de cada página. Este modelo, a pesar de que ha sido útil en su aplicación, tiene tres inconvenientes: el primero es que para reconocer las regiones de forma automática se necesita el mismo tiempo de la transcripción de una página (doblando el tiempo total); además, el recurso a este modelo consume también créditos de uso del *software*; por último, no siempre funciona perfectamente y, en algunas páginas, se necesita igualmente intervenir manualmente, o escribir en el archivo TXT los símbolos de puntuación que quedan fuera de las regiones en el margen derecho.

CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

5. Principales dificultades de transcripción y toma de decisiones

Tomando en consideración las características tipográficas y metalexicográficas de la disposición del texto en las páginas y de la macroestructura y microestructura antes descritas, se van a ilustrar aquí las principales dificultades y las decisiones que se han tomado para resolver algunos problemas surgidos con la transcripción.

5.1 *Por lo que concierne a la disposición del texto*

El primer obstáculo atañe a la división de la página en tres columnas. Como se ha descrito más arriba, fue necesaria la segmentación del texto en tres regiones, previa a la transcripción. Si se solicitaba al *software* directamente el reconocimiento del texto sin crear las regiones, el programa leía las líneas de las tres columnas seguidas, como si se tratara de una sola y de un único párrafo, de izquierda a derecha. Por esta razón, cuando no se aplicaba la transcripción automática con el modelo *layout*, antes de empezar la transcripción de cada página había que intervenir manualmente con la herramienta *Region* de *Transkribus* y crear las tres regiones. Con ello, se conseguía dejar fuera de las regiones los elementos que no se deseaba transcribir, como, por ejemplo, el número de las páginas, las tres letras-guía encima de cada columna o las signaturas tipográficas de numeración del tomo o de los pliegos. Los resultados obtenidos ya en el primer archivo fueron satisfactorios: después de la división en regiones de forma manual, el programa leía separadamente las líneas de cada región, aunque a veces fallaba y era necesario segmentar las líneas a mano. Este fallo esporádico se ha mantenido hasta el final del trabajo.

5.2 *Por lo que concierne a la microestructura*

La microestructura del lemario también proporcionó bastantes retos durante el reconocimiento del texto y la transcripción. Por lo que atañe al símbolo de la doble daga, el problema se resolvió rápida y fácilmente. Al principio, el programa leía este símbolo con resultados de transcripción diferentes (a veces con letras, como una < t > minúscula, por ejemplo, o bien con un corchete). Se decidió intervenir manualmente en las primeras páginas, convirtiendo la doble daga en dos barras inclinadas (//) y el programa aprendió muy rápidamente a sustituir este símbolo. En cambio, la llaves y los corchetes han representado problemas que no

se han conseguido solucionar.

Como puede imaginarse, la llave no se puede leer en una sola línea (ya que recoge los lemas de al menos dos líneas, si no más). El programa detectaba la sinuosidad de la parte superior del símbolo de la llave y la transcribía como una C o una S; además, no repetía los equivalentes de traducción que comparten esos lemas. En general, las propuestas de reconocimiento del símbolo eran variadas y confusas y, por ello, hubo que tomar una decisión al respecto, interviniendo cada vez manualmente. Se ha adoptado la convención de transcribir, seguidas en serie, todas las entradas englobadas dentro de una llave con sendas iniciales mayúsculas y separadas por una coma (o por un punto si antes hay una abreviación de la categoría gramatical), transformándolas así en una única entrada con más lemas; siguiendo luego con la serie microestructural básica de las demás entradas (categoría gramatical y equivalente de traducción), como puede verse en las imágenes 7 (a, b) y 8 (a, b), en las que se muestra el texto original y la transcripción manual en *Transkribus* de dos ejemplos de uso de la llave.

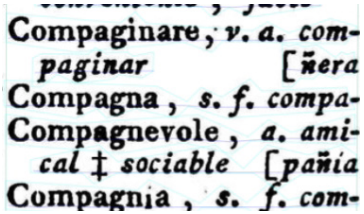
Busbaccheria, *s. f. engaño* Busbacco, Busbaccone, *s. m. engañador* Busca, *s. f. busca*	Busbaccheria, s. f. engaño Busbacco, Busbaccone, s. m. engañador Busca, s. f. busca
Imagen 7a	Imagen 7b

Cocimento, Cocitura, *s. m. s. f.* *escozor ‡ coccion ‡ quemadura*	Cocimento, s. m. Cocitura, s. f. escozor // coccion // quemadura
Imagen 8a	Imagen 8b

Por lo que atañe al corchete, en cambio, la dificultad se ha podido solucionar parcialmente. El programa detectaba, exactamente como estaba escrito, el corchete de apertura en la línea de arriba, siguiendo al equivalente de traducción de la entrada anterior. Pensamos que la presencia de este corchete podía causar problemas a la hora de convertir el TXT en formato digital y dar instrucciones al programa de digitalización. Por esta razón, se ha preferido borrar totalmente el contenido del corchete y añadirlo manualmente donde correspondía, es decir, en la línea de abajo (véanse imágenes 9a y 9b). Después de varias operaciones de eliminación del contenido del corchete, el programa aprendió a no leer esta parte de línea.

CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

Obviamente, la operación de completar la línea de abajo se ha tenido que realizar siempre manualmente.

	Compagna, s. f. compañera Compagnevole, a. ami- cal // sociable Compagnia, s. f. compañía
Imagen 9a	Imagen 9b

Por lo que respecta, en cambio, al guion al final de la línea que indica que la palabra cortada sigue en la línea siguiente (véase *amical* en la imagen 9a y su transcripción en 9b), se ha tomado una primera decisión, la cual consiste en mantener dicha división para obtener un archivo de texto TXT lo más fiel posible al original −para posibles futuros usos del texto, como podría ser una edición filológica crítica−, ya que se está a tiempo de tomar una segunda decisión por la que en el archivo TXT se elimina el guion seguido de un salto de línea, juntando así las palabras cortadas.

6. Detección y corrección de erratas

El control de la transcripción automática del texto del tomo I ha identificado una serie de erratas y de errores[16]. Del análisis ecdótico de los errores (conceptuales) y de las decisiones que se han tenido que tomar al respecto se proporcionará algún ejemplo en §7. Nos concentraremos en este párrafo en las erratas, anticipando que no ha sido siempre posible reconocer e interpretar los fallos para intervenir con convicción. Como se ha especificado más arriba, para poder tomar una decisión ante una posible errata se han cotejado los cuatro ejemplares en PDF a disposición del lemario italiano-español del diccionario de Cormon y Manni de 1805: BUCM, BNF, BNE y BMF.

16 Adoptamos esta distinción terminológica, considerando *errata* los fallos de ejecución o errores materiales; es decir, una equivocación humana, del autor o del cajista. Recurriremos, en cambio, a la denominación de *error* para clasificar los fallos de tipo cognitivo, que derivan de la falta de conocimiento del autor, del impresor, tipógrafo, cajista, etc. La errata es material y el error, conceptual (Fernández-Quesada, Rodríguez-Rubio 2022).

6.1 *Erratas de fallos mecánicos*

Las erratas que hemos identificado en la obra son de diferentes categorías, siendo preponderantes los fallos mecánicos o de distracción que se listan a continuación. Se trata, por proporcionar solamente uno o pocos ejemplos de cada caso, de fallos como los siguientes:

- Inversión de tipos: algunas palabras presentan dos tipos posicionados de manera invertida con respecto a la palabra correcta, como sucede con "ciudado"[17] en vez de "cuidado".
- Sustitución de tipos: algunas letras se han sustituido erróneamente con otras, como en "palebras" donde debía decir "palabras", "antivedimenro" por "antivedimento".
- Omisión de tipos: cuando algunas palabras carecen de una letra. Es lo que sucede con "cabeller_" (por "cabellera") o en "asolv_miento" por "asolvimiento", "cub_ir de moho" en vez de "cubrir".
- Colocación de un tipo al revés: se ha encontrado palabras que contienen un tipo girado (derecha/izquierda o bocabajo); el caso más repetido es el de las letras < u > y < n > ("Dnbbioso" en vez de "Dubbioso").
- Letras mayúsculas fuera de lugar: aparecen letras mayúsculas en medio de algunas palabras y lemas, como en "BAldo".
- Adición de sílabas: se trata de palabras que contienen una sílaba de más, como en "dañadado" (en vez de "dañado") o que las tienen repetidas ("impepenitencia").
- Inversión de sílabas: "peresozo" en vez de "perezoso".
- Errores en las abreviaturas de las categorías gramaticales (que no se han hallado en diccionarios que pudieron ser fuente de este): por ejemplo, "s. f." para "Ateo" en vez de "s. m.", debido, probablemente, al hecho de que se ha unido con una llave al lema "Ateista".

Estos fallos se han considerado totalmente mecánicos, de distracción o descuido. La primera decisión que se tomó al respecto −salvo en casos como el de los tipos colocados al revés− fue la de no intervenir, pues el primer objetivo de la investigación y del proceso de reconocimiento y transcripción automática era educar al *software* para que aprendiese a leer y transcribir el texto original lo más fielmente posible. De cada una de estas erratas y decisiones se tomó oportuna nota en un diario de trabajo para no perder la huella de los cambios realizados,

17 Los subrayados para resaltar el error son siempre nuestros.

CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

poder localizarlas más adelante y corregirlas en el archivo TXT que será el texto base de la codificación.

6.2 *Erratas en el uso de las tildes*

No va a ser posible detenerse en este estudio sobre cuestiones ortográficas en sentido amplio; incluso remitir a estudios fundamentales de historiografía ortográfica ocuparía excesivo espacio[18]. Se presentará aquí un breve listado de ejemplos de erratas en el uso de las tildes que no nos parece que correspondan a la ausencia o caos normativo de aquella época. La variedad de imprecisiones encontradas en la acentuación es evidente y no ha resultado siempre fácil tomar una decisión. Entender si la adición u omisión de una tilde se debe a la distracción del tipógrafo, a un desconocimiento sobre la lengua extranjera con la que se estaba editando, a una diferente y efímera regla de acentuación que se estaba aplicando en ese momento o a una mancha de tinta ha sido un trabajo arduo. Este tipo de análisis −con continuos cotejos de los ejemplares a disposición, del *Nuevo tesoro lexicográfico de la lengua española*, de las cinco ediciones del *Vocabolario* de la Crusca en línea y del corpus CORDE− puede llevar a convertir una errata en un error de valor ecdótico, para el que hay que aplicar un tratamiento diferente. Siendo copiosas las erratas de acentuación, listaremos solo algún ejemplo de los diferentes tipos más difundidos.

- Tildes en vocales no tónicas: “anímal”, “provídencia”, “arbóles”, “claúsula”, “fástidioso”, “fuligínoso”, “márfil”, etc.
- Ausencia de tildes: “montañes”.
- Tildes en letras de abreviaturas: “Dolcione, á.” (de adjetivo).
- Tildes graves en palabras escritas en español: “jugar à”, “cèdula”, “de à dos pies”.
- Tildes agudas en palabras del italiano que no hemos conseguido documentar que se escribieran en aquel periodo también así: “giovedí”, “cosí”.
- Uso del acento circunflejo en palabras que no contienen la <x> con el sonido del grupo consonántico latino [ks] o falta de uniformidad en la aplicación de esta norma: “exâmen” frente a “exámen”, “divulgacîon”.

18 Es impensable poder indicar aquí todas nuestras referencias bibliográficas, estudios historiográficos sobre la ortografía, ni siquiera limitándonos al periodo del siglo XIX. Por lo que se refiere a la aplicación de las normas de ortografía en la lexicografía de principios del XIX, vamos a remitir solo a Terrón Vinagre (2022) y a Blanco Izquierdo (2021); por lo que se refiere a la ortografía del español en los bilingües de italiano-español, a De Hériz (2016, 2018).

7. El cotejo de fuentes y ediciones posteriores de la obra a favor de la transcripción y digitalización

Cuando se revisa la transcripción del texto original en formato .txt y se interviene en la corrección, se observa a menudo defectos de nitidez tipográfica que se pueden resolver, a veces, con el cotejo de los ejemplares de la misma edición a disposición; otras erratas como las ejemplificadas en §6 no se pueden interpretar sino con la deducción lógica o analógica. Por otro lado, también se van identificando erratas o errores que se repiten en todos los lemarios consultables y que por diferentes razones sorprenden más que las anteriores y obligan a tomar decisiones no siempre fáciles de adoptar. Si la transcripción de un original (automática o manual) tiene como único objetivo reproducir un texto antiguo en un formato de fácil lectura o bien en una edición crítica absolutamente fiel al original, no se suele intervenir sobre esos defectos considerados errores, sino que se remite a notas críticas que interpretan la intención del autor o señalan correcciones adoptadas en las siguientes ediciones o en textos de autores que tomaron como fuente el fragmento en cuestión.

Dado que los objetivos de nuestro proyecto no se limitan a generar protocolos de transcripción automática, a experimentar solamente programas informáticos de transcripción automática, sino que prevén el paso siguiente de la codificación del texto transcrito para hacerlo accesible junto con otros de la misma categoría lexicográfica bilingüe con el fin de constituir un tesoro lexicográfico digital consultable, en algunas ocasiones ha sido necesario plantearse cuál era la intención metalexicográfica del autor o autores del tomo I del Cormon y Manni de 1805 respecto a algunas peculiaridades de la macro y microestructura de esta versión. Claro está que estas extrañezas también pudieron ser fruto de los tipógrafos, quienes, ante una duda, consultaron quizás otros diccionarios para resolverla y no ralentizar el proceso de impresión de la obra. A continuación, se proporcionarán algunos ejemplos de resolución de problemas mediante el cotejo de obras lexicográficas anteriores y posteriores a la de Cormon y Manni italiano-español de 1805.

7.1 *Cotejo de la primera (1805) y segunda (1821) edición del Cormon y Manni italiano-español*

Ante todo, cabe confirmar la convicción de Martínez Egido (2010: 71) al respecto, el cual, tras haber cotejado una cala y los paratextos de la 2ª edición de 1821

CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

del diccionario con la 1ª de 1805, aseguraba que el *Diccionario italiano-español y español-italiano* de 1821 de Cormon y Manni no contiene más diferencias que la ausencia en el título de la portada de la caracterización "de faltriquera" (tomo I) y "da tasca" (tomo II). Esto se puede corroborar ya que se han consultado varios ejemplares de 1821 para cada uno de los errores, caracteres poco nítidos, manchas extrañas, signaturas tipográficas, etc., presentes en los cuatro ejemplares de 1805 y no se ha observado la más mínima diferencia; solo la apenas mencionada del título de la obra.

7.2 *Erratas, errores y curiosidades de la macroestructura*

Tal como se ha descrito más arriba (§3), la macroestructura de este lemario se caracteriza por el uso de las llaves para englobar variantes morfológicas de una palabra-lema que correspondan al mismo equivalente de traducción (o a más de uno y distintas acepciones de significado). A continuación presentamos dos ejemplos de lematización repetida (en *Apparato* y *Malaventura*). Propondremos como ejemplo de cotejo −aunque nos consten más− el C&M/ita-spa-1805[19] con otro diccionario bilingüe de 1802 de los mismos autores (Cormon y Manni) para la combinación italiano-francés, el *Dictionnaire portatif et de prononciation, italien-français et français-italien*, publicado también en Lyon por parte de los mismos libreros Cormon y Blanc (Lillo 2019: 219-223; Mormile 1993: 67-71). Alvar Ezquerra (2010: 48-9) ya supuso una probable filiación del bilingüe italiano-español de 1805 respecto al italiano-francés de 1802 de los mismos autores.

19 A partir de ahora y para abreviar al comentar cotejos entre diferentes obras, citaremos con la sigla C&M/ita-spa-1805 el tomo I del lemario italiano-español del *Diccionario de faltriquera* de Cormon y Manni de 1805 y con la sigla C&M/ita-fra-1802 el lemario italiano-francés del diccionario de Cormon y Manni de 1802.

Apparato, s.m. } aparato, preparativo ‡ Apparato, s.m. } adorno, decoracion	*Apparato*, ta, adj. Appris. + Préparé, etc. + Fourni ; pourvu. *Apparato*, subst. m. (*ap-pa-ra-to*) Appareil; apprêt; préparatif. + Apparat. + *Apparato di chiesa.* Ornement ; décoration ; pompe ; parure. + *Apparato militare*, apprêt, appareil militaire.
Malaventura, } s.f. mala-ventura Malaventura, }	*Malaventura*, et *Mala ventura*, s. f. (*-ven-toù-ra*) Malheur ; disgrace ; malencontre.
C&M-ita-spa-1805	C&M-ita-fra-1802

Apparato en C&M/ita-spa-1805 está etiquetado con una sola categoría gramatical (s. m.). En otros diccionarios anteriores se lematiza con dos entradas homónimas: una para el adjetivo derivado de *apparare* y la otra para el sustantivo masculino con varias acepciones[20]. Parece evidente que se puede considerar esta repetición del lema con la misma abreviatura gramatical como un error de distracción; por lo tanto, en la fase de digitalización de la entrada *apparato* podremos decidir que la 1ª forma se categorice como un adjetivo y la 2ª mantenga la abreviatura del original (s.m.). Conviene, sin embargo, ser conscientes de que una intervención de este tipo obliga a plantearse y a decidir si es necesario añadir a la forma del primer lema la desinencia femenina (*Apparato*, *ta*). Algo similar sucede con el lema compuesto *Malaventura* y su variante gráfica *Mala ventura*; se trata de un error que habrá que resolver durante la codificación para la digitalización del texto.

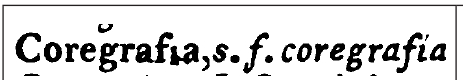
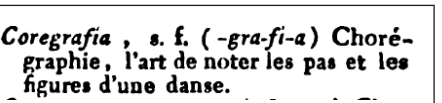
En otras ocasiones, algunas erratas debidas aparentemente a la confusión de tipos llevan a la más que probable identificación de la fuente, la cual contiene un verdadero error. Así hemos interpretado la forma del lema *Coregrafia*, no documentada en ninguna otra fuente lexicográfica monolingüe del español y del italiano[21] y sí, únicamente, en el bilingüe C&M-ita-fra-1802. Se trata probablemente de un galicismo del momento; el sonido del francés *Chorégraphie* se calcó en una

20 Ya en la 3ª edición del *Vocabolario degli accademici della Crusca* (1691). En estos cotejos daremos muestra de semejanzas o diferencias con diccionarios con fecha de edición más cercana a la del nuestro.

21 No se debe confundir este lema con el de *Corografia* (lematizado y traducido correctamente en los dos diccionarios que se están cotejando). Por otro lado, *Coreografia* no está recogido en el *Vocabolario* de la Crusca hasta la 5ª edición (1863-1923) y, según el *NTLLE*, hasta 1825 en español en el *Diccionario* de Núñez de Taboada. Hemos encontrado en Gherardini (1840) un comentario con el que explica por qué recurre en un diccionario a la marca diatécnica *Coregrafia* y no a *Coreografia* (1840: 766).

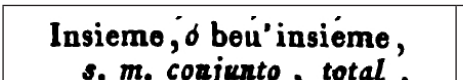
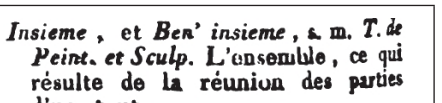
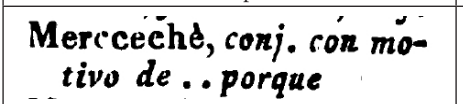
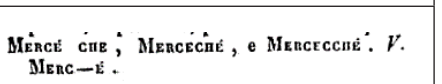
CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

inexistente forma italiana y se trasladó a otra española que tampoco encontramos documentada en el *NTLLE*.

	
C&M-ita-spa-1805	C&M-ita-fra-1802

En este caso, habrá que decidir si se digitaliza y codifica tal cual la entrada en el Tesoro TELEI, manteniendo presente en el diccionario en línea una "palabra fantasma" o de muy breve vida (Álvarez de Miranda 2000: 56-73)[22].

No hemos hallado siempre la huella de las erratas de colocación de tipos en las palabras-lema. Aun así, el cotejo con otros diccionarios anteriores al C&M-ita-spa-1805 –o posteriores, pero cercanos a la fecha de edición– han avalado la toma de una decisión respecto a una corrección en la transcripción y a la forma que se digitalizará. Véanse estos dos ejemplos:

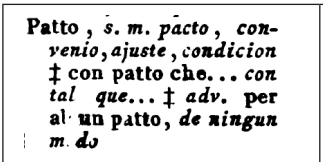
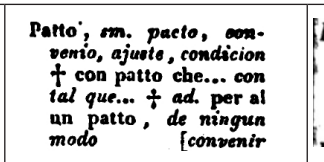
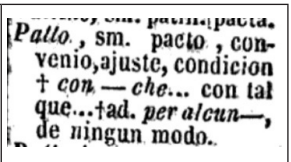
	
C&M-ita-spa-1805	C&M-ita-fra-1802
	
C&M-ita-spa-1805	Vanzon 1841

Es lógico –pero podemos basarlo en modelos lexicográficos– que la segunda forma *Beu'insieme* de la entrada *Insieme* se debe corregir como *Ben' insieme*. Asimismo nos parece coherente que la intención de lematizar la conjunción *Mercecché* con dos <cc> se respete con una forma correcta: o sea, *Mercecché* (en vez de *Mercceché*).

22 El error perdura en el C&M italiano-español de 1821 y Blanc en su revisión de la obra de 1843 no lo elimina. Salvo que se consiga demostrar el uso de este término en ese periodo, este lema y su equivalente bien sirven de ejemplo de las palabras fantasma de la lexicografía.

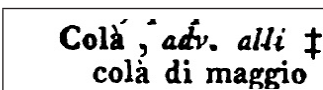
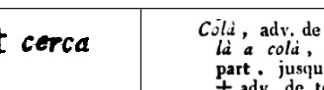
7.3 Erratas, errores, imperfecciones y curiosidades de la microestructura

Los problemas o dudas surgidos durante la transcripción de los contenidos de los artículos también se han intentado resolver con el cotejo de otras obras lexicográficas. Como se ha explicado en §3 la imprenta en prensa con tipos y tinta producía muchas manchas o falta de nitidez en la impresión. Para completar algunas palabras en las que no se leen bien los caracteres, la comparación entre la 1ª edición (1805) y la 2ª (1821) no ha resuelto nunca las dudas, pues las galeras parecen ser exactas. En cambio, ha sido muy útil poder comparar dichos artículos con los de la edición de 1843 o 1848 del Cormon y Manni italiano-español, la revisada por S. H. Blanc (Alvar Ezquerra 2010: 52-5). Como muestra de muchas comparaciones realizadas presentamos el artículo del lema *Patto*, en el que dos palabras de la penúltima y última línea están incompletas:

 Patto , *s. m. pacto , convenio, ajuste , condicion* ‡ con patto che... *con tal que...* ‡ *adv.* per al un patto, *de ningun m. do*	 Patto', *sm. pacto, convenio, ajuste, condicion* † con patto che... *con tal que...* † *ad.* per al un patto , *de ningun modo* [*convenir*	 [*Patto* , sm. pacto , convenio,ajuste, condicion † *con — che...* con tal que...†ad. *per alcun—*, de ningun modo.
C&M-ita-spa-1805	C&M rev. Blanc-ita-spa-1843	C&M-ita-spa-1848

En las tres muestras propuestas, la microestructura de las tres ediciones no es idéntica por lo que a símbolos metalexicográficos se refiere, pero con este doble cotejo se completan con seguridad "alcun patto" y "de ningun modo" de la edición de 1805.

Cuando nos sorprende la ausencia de equivalentes en español dentro del artículo o la traducción de ejemplos, también consultamos diccionarios anteriores y el Cormon y Manni de italiano-francés de 1802. De esta manera solemos entender cuál era el proyecto del contenido de la entrada, si bien en ningún caso hemos añadido todo lo que probablemente falta. Es lo que sucede en el artículo del lema *Colà*, comparado aquí abajo, donde no se distinguen las dos funciones adverbiales y no se traduce el ejemplo *colà di maggio*.

 **Colà , *adv. alli* ‡ *cerca* colà di maggio**	 *Colà* , adv. de lieu. (*ko-là*) Là. + *Di là a colà* , de l'endroit d'où l'on part , jusqu'où l'on veut se rendre. + adv. de temps. Environ ; vers. *Colà di maggio.*
C&M-ita-spa-1805	C&M-ita-fra-1802

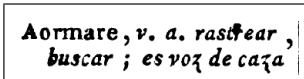
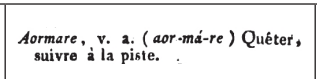
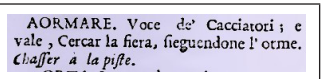
CUADERNOS AISPI 26 (2025): 45-71
ISSN 2283-981X

También ha sucedido que nos ha extrañado la ausencia de algún símbolo de valor metalexicográfico, como la doble daga que antecede a las diferentes acepciones de significado. Presentamos aquí el ejemplo del artículo del lema *Manco*, en el que, al principio, nos llamó la atención la ausencia de una coma entre "menos" (español) y "venir manco" (en italiano). La edición del Cormon y Manni revisada por Blanc (1843) no resolvía la duda respecto al elemento que faltaba, mientras que el cotejo con el diccionario italiano-español del diccionario anónimo publicado en 1853 por la librería parisina de Rosa y Bouret (Castillo Peña 2010)[23] nos ha permitido entender que entre el equivalente "ménos" y la colocación "venir manco" falta en C&M de 1805 el símbolo de acepción (doble daga en C&M 1805), que podremos introducir en la digitalización del artículo.

Manco, *ad. ménos* venir manco, *faltar* ; *etc* ‡ senza manco, *sin falta* ‡ non ci ho manco	mancamiento † *ad. menos* venir manco, *faltar, etc.* † senza manco, *sin falta* † non ci ho manco pensato, *ni siquiera he pensado á ello* [miento	MANCO, *ad.* Menos. | *Venir manco*; faltar. | *Senza manco*; sin falta. | *Non ci ho manco pensato*; ni siquiera he pensado en ello.
C&M-ita-spa-1805	C&M rev Blanc-ita-spa-1843	R&B-ita-spa-1853

Consideramos oportuno concluir esta sección con un ejemplo útil para demostrar que el Cormon y Manni italiano-francés (1802), aunque se pueda confirmar que fue la fuente principal del lemario italiano-español (1805) de los mismos autores, no constituyó la única base de inspiración del quehacer lexicográfico. El elemento microestructural que aportamos como botón de muestra para apoyar esta tesis es una glosa diatécnica añadida al final del artículo del lema *Aormare*: "es voz de caza", información que no se encuentra en el lemario italiano-francés, como puede verse abajo. La Crusca lematiza *Aormare* a partir de la 3ª edición de su *Vocabolario* (1691) y marca este uso relacionado con la caza tanto en esa edición como en la 4ª (1729-1738), ambas anteriores a nuestro diccionario. Sugerimos aquí una comparación con el artículo registrado en el *Dictionaire* italiano, latín y francés de Antonini (ya mencionado en §3 a propósito del uso de las llaves para unir formas de lemas).

23 Carmen Castillo Peña (2010: 154-68) ha demostrado con un largo estudio de cotejos macro y microestructurales que el Cormon y Manni italiano-español/español-italiano de 1805 es una de las fuentes principales (no la única) del lemario italiano-español del anónimo *Nuevo diccionario italiano-español* de los editores parisinos Rosa y Bouret (1853) y no lo son, en cambio, las ediciones del Cormon y Manni de 1843 y 1848 revisadas por Blanc.

		
C&M-ita-spa-1805	C&M-ita-fra-1802	Antonini-ita-fra-1743

8. Conclusiones

La transcripción automática de un diccionario bilingüe impreso a principios del siglo XIX, como la del lemario italiano-español del Cormon y Manni de 1805 que se ha descrito, puede parecer *a priori* un desafío más simple de afrontar que el que representarían obras de siglos anteriores. Se ha visto, en cambio, que cada una de estas obras conlleva diferentes dificultades por las condiciones materiales de los ejemplares, por la peculiaridad tipográfica de los textos de los lemarios y por las muchas erratas y errores que se van encontrando en la revisión de las 422 páginas del texto transcrito automáticamente.

Dado que la transcripción de dicho texto no ha sido el objetivo final del proyecto de investigación, sino el paso previo a la digitalización con la que se codificará cada uno de los elementos macro y microestructurales, cabe recordar que los ejemplos que se han aportado en esta publicación son un botón de muestra de todas las consideraciones que hubo que plantearse durante la revisión del texto transcrito; consideraciones intra y extratextuales.

Bibliografía citada

Alvar Ezquerra, Manuel (2010), "Un siglo de lexicografía bilingüe español-italiano", *Diversidad lingüística y diccionario,* eds. Marta Concepción Ayala; Antonia María Medina. Málaga, Universidad de Málaga: 45-118.

Álvarez de Miranda, Pedro (2007), "Palabras y acepciones fantasma en los diccionarios de la Academia", Alicante, *Biblioteca virtual Miguel de Cervantes* [20/03/2025] <https://www.cervantesvirtual.com/nd/ark:/59851/bmc2f836>.

Anónimo (1853), *Nuevo diccionario italiano-español*, Paris, Librería de Rosa y Bouret.

Antonini (1743), *Dictionaire italien, latin e françois*, Paris, Prault Fils.

Blanco Izquierdo, M.ª Ángeles (2021), "Más allá de las letras: la acentuación gráfica en el *DRAE* (1869, 1884 y 1899)", *El diccionario académico en la segunda mitad del siglo XIX:*

evolución y revolución, eds. María Ángeles Blanco; Gloria Clavería. Berlin, Peter Lang: 91-125.

Bohbot, Hervé; Frontini, Francesca; Luxardo, Giancarlo; Khemakhem, Mohamed; Romary, Laurent (2018), "Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré", *GLOBALEX 2018,* May 2018. Miyazaki, Japan: 1-6 [7/3/25] <https://hal.science/hal-01728328v1>.

Bruña Cuevas, Manuel (2008), "La producción lexicográfica con el español y el francés durante los siglos XVI al XIX", *Philologia Hispalensis*, 22: 37-111.

Castillo Peña, Carmen (2010), "El *Nuevo diccionario italiano-español* (1853) de los editores Rosa y Bouret", *Textos fundamentales de la lexicografía italoespañola: (1805-1916)*, dir. Félix San Vicente. Monza, Polimetrica: 146-92.

Castillo Peña, Carmen (2020), "Epigrama: Un portal para la edición digital de textos gramaticales", *Anales De Lingüística*, 4: 201-217 [20/03/2025] <https://revistas.uncu.edu.ar/ojs3/index.php/analeslinguistica/article/view/4395>.

Cazorla Vivas, M.ª del Carmen (2002), *Lexicografía bilingüe de los siglos XVIII y XIX con el español y el francés*, Tesis doctoral, Madrid, Universidad Complutense.

Cazorla Vivas, M.ª del Carmen (2010), "Panorama de la lexicografía bilingüe y plurilingüe del español a comienzos del siglo XIX", *Textos fundamentales de la lexicografía italoespañola (1805-1916),* dir. Félix San Vicente. Monza, Polimetrica: 27-56.

Cormon, J. L. B.; Manni, Vincenzo (1802), *Dictionnaire portatif et de prononciation italien-français et français-italien*, Lyon, Libraires B. Cormon et Blanc.

Cormon, J. L. B.; Manni, Vincenzo (1805). *Diccionario de faltriquera italiano-español y español-italiano*, Leon, Librería de B. Cormon y Blanc.

Cormon, J. L. B.; Manni, Vincenzo (1821), *Diccionario italiano-español y español-italiano*, Leon, Librería de B. Cormon y Blanc.

Cormon, J. L. B.; Manni, Vincenzo (1843), *Diccionario italiano-español y español-italiano*, Nueva edición revista y aumentada por S. H. Blanc, Leon, Librería Cormon y Blanc.

Cormon, J. L. B.; Manni, Vincenzo (1848), *Diccionario italiano-español y español-italiano*. Nueva edición revista y aumentada por S. H. Blanc, Leon y Paris, Librería St. Hilaire Blanc y C^{ia}.

Costa, Rute; Salgado, Ana; Khan, Anas Fahad; Carvalho, Sara; Romary, Laurent *et al.* (2021) "MORDigital: The Advent of a New Lexicographical Portuguese Project", *eLex 2021 - Seventh biennial conference on electronic lexicography*, Jul 2021, Brno, Czech Republic [20/03/2025] <https://inria.hal.science/hal-03195362v2>.

De Hériz, Ana Lourdes (2016), "Normas ortográficas de la lengua española en la lexicografía ítalo-española del siglo XIX", *Orillas*, 5 [30/03/2025] <https://www.orillas.net/orillas/index.php/orillas/article/view/355/350>.

De Hériz, Ana Lourdes (2018), "Las reformas de la ortografía en los diccionarios bilingües de italiano-español de los siglos XIX, XX y XXI", *Boletín de la Sociedad Española de la Historiografía Lingüistica*, 12: 95-121 [30/03/2025] <http://www.sehl.es/010_05_ana-lourdes_de-heacuteriz.html>.

Fernández-Quesada, Nuria; Rodríguez-Rubio, Santiago (2022), "El tratamiento del error textual y de la errata en la era digital: elogio de la corrección", *Detección y tratamiento de errores y erratas: un diagnóstico para el siglo XXI*, eds. Nuria Fernández-Quesada, Santiago Rodríguez-Rubio. Madrid, Dykinson: 13-28.

Gallina, Anna Maria (1991). "La lexicographie bilingue espagnol-italien, italien-espagnol". eds. Franz Josef J. Hausmann *et al. Wörterbucher Dictionaries Dictionnaires*. Berlin-New York, Walter de Gruyter: 2991-97.

Gattel, Claude Marie (1798), *Nuevo diccionario portatil español y francés,* Paris, Casa de Bossange, Masson y Besson.

Gattel, Claude Marie (1803), *Nuevo diccionario portatil español é inglés*, Valencia, P. J. Mallen y C.

Gherardini, Giovanni (1840), *Voci e maniere di dire italiane additate a'futuri vocabolaristi*, Volume II, Milano, Gio. Bat. Bianchi di Giac.

Khemakhem, Mohamed; Foppiano, Luca; Romary, Laurent (2017), "Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields", *Electronic Lexicography*, eLex 2017, Sep 2017, Leiden, Netherlands [07/3/2025] <https://hal.science/hal-01508868v2>.

Khemakhem, Mohamed; Galleron, Ioana; Williams, Geoffrey; Romary, Laurent; Ortiz Suárez, Pedro Javier (2019), "How OCR Performance can Impact on the Automatic Extraction of Dictionary Content Structures", *19th annual Conference and Members' Meeting of the Text Encoding Initiative Consortium (TEI) -What is text, really? TEI and beyond*, Sep 2019, Graz, Austria [07/03/2025] <https://hal.science/hal-02263276v1>

Lillo, Jacqueline, ed. (2019), *1583-2010: Quattro secoli e più di lessicografia italo-francese. Repertorio analitico di dizionari bilingue*, Vol I, 2ª edizione riveduta e ampliata, *Quaderni del Cirsil, 14*. Bologna, Università di Bologna.

Lindemann, David; Alonso, Mikel (2021), "A workflow for historical dictionary digitization: Larramendi's Trilingual Dictionary", *Proceedings of eLex,* 598-614 [07/03/2025] <https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_39_pp598-614.pdf>.

Lindemann, David; Khemakhem, Mohamed; Romary, Laurent (2018), "Retro-digitizing and Automatically Structuring a Large Bibliography Collection", *European Association for Digital Humanities (EADH) Conference*, EADH, Dec 2018, Galway, Ireland [07/03/2025] <https://hal.science/hal-01941534v1>

Lombardero Caparrós, Alberto (2015), *The Historiography of English Language Teaching in Spain: A Corpus of Grammars and Dictionaries (1769-1900)*, Tarragona, Universitat Rovira i Virgili.

Martínez Egido, José Joaquín (2008), "Origen y desarrollo positivo de la lexicografía bilingüe español-italiano (siglos XVI-XIX)", *Philologia hispalensis*, 22: 213-58.

Martínez Egido, José Joaquín (2010), "El *Diccionario de faltriquera italiano-español y español-italiano* de J. L. B. Cormon y V. Manni (1805)", *Textos fundamentales de la lexicografía italoespañola (1805-1916)*, dir. Félix San Vicente, Monza, Polimetrica: 57-92.

Mormile, Mario (1993), *Storia dei dizionari bilingui italo-francesi. La lessicografia italo-francese dalle origini al 1900*, Fasano, Schena Editore.
Nalesso, Giulia (2024), "Teoría y práctica de la retrodigitalización de diccionarios: el caso del *Vocabulario de las dos lenguas toscana y castellana*", *Artifara*, 24.1: 135-53 [20/03/2025] <https://ojs.unito.it/index.php/artifara/article/view/9988/8605>
Núñez de Taboada, Melchor Manuel (1825), *Diccionario de la lengua castellana*, Paris, Seguin.
Rabus, Achim (2019), "Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus", *Scripta & e-Scripta,* 19: 9-32.
San Vicente Santiago, Félix, dir. (2010), *Textos fundamentales de la lexicografía italoespañola: (1805-1916)*. Monza, Polimetrica.
San Vicente, Félix, dir., *Contrastiva. Portal de gramática y de lingüística contrastiva español-italiano*. [20/03/2025] < http://www.contrastiva.it>.
Terrón Vinagre, Natalia (2022), *Lexicografía y ortografía en el siglo XIX*, Berlin, Peter Lang.
Vanzon, Carlo Antonio (1841), *Dizionario universale della lingua italiana*, Palermo, Tipografia Demetrio Barcellona.
Wilkinson, Mark, *et al.* (2016), "The FAIR guiding principles for scientific data management and stewardship", *Sci. Data*, 3:160018 [7/3/2025] <https://www.nature.com/articles/sdata201618>.

Florencia Ferrante es licenciada en Letras por la Universidad de Buenos Aires y en Italianística por la Universidad de Bolonia. En 2019 se doctoró en la Universidad de Módena y Reggio Emilia con una tesis sobre el pensamiento crítico de Juan Rodolfo Wilcock. Se ha ocupado principalmente de las relaciones literarias e intelectuales entre Italia e Hispanoamérica, especialmente en el ámbito de la teoría literaria y de la historia de la traducción. Es autora de una monografía (*Juan Rodolfo Wilcock critico*, ETS, 2022) y de varios artículos publicados en revistas como *Strumenti Critici*, *Hispamérica* y *Meta*.
florencia.ferrante@unige.it

Chiara Valente es licenciada en Lenguas y culturas modernas por la Universidad de Génova. De 2018 a 2021 formó parte del grupo de investigación GramHisGram de la Universidad de Salamanca, donde también defendió su tesis doctoral en 2021 con un estudio sincrónico, diatópico y comparativo español-italiano del uso de las formas verbales del PPS y PPC. Actualmente es profesora de lengua y literatura española en la *Scuola Secondaria di II grado* y profesora *a contratto* de la Universidad de Génova. Es autora de capítulos de libros en editoriales como Dykinson, De Gruyter y Arco/Libros y de varios artículos en revistas como *Moenia* y *Lingua e Stile*.
chiara.valente@edu.unige.it

Ana Lourdes de Hériz es profesora titular de Lengua española en la Universidad de Génova. Se doctoró en la Universidad de Pisa con una tesis doctoral sobre la colaboración en la redacción de heterobiografías, estudio realizado con un enfoque pragmático. Durante los años en los que trabajó en la universidad como lectora investigó sobre el análisis del error, dedicándose luego a la historia de la traducción al español en España en el siglo XIX, a la gramática contrastiva (español-italiano) y, sobre todo, a la lexicografia en el siglo XIX y a la metalexicografía actual. Ha coordinado dos grupos de investigación PRIN (2017 y 2022) y publicado capítulos de libros en editoriales italianas y europeas.
ana.deheriz@unige.it