

# MARIA FELISA BERMEJO CALLEJA GIULIA LANTERI MARIÁ JOAQUINA VALERO GISBERT MARIA CARMELA ZACCONE LA TRANSCRIPCIÓN DIGITAL DEL DICCIONARIO DE L. AMBRUZZI

Università degli Studi di Torino  
Università degli Studi di Parma

## Resumen

El *Nuovo dizionario* de Lucio Ambruzzi (DBA), publicado en 1948-1949, es una obra bilingüe con elementos enciclopédicos. Su digitalización enfrenta desafíos por su compleja microestructura y abundancia de abreviaturas. Se busca sistematizar su ordenación, clarificar abreviaturas y mejorar su consulta. El estudio aborda aspectos metalexicográficos y lingüísticos para optimizar su uso en formato digital.

palabras clave: lexicografía bilingüe, digitalización, abreviaturas

## Abstract

### *The digital transcription of the dictionary by L. Ambruzzi*

Lucio Ambruzzi's bilingual dictionary (DBA), published between 1948 and 1949, is a work that features encyclopedic elements. Its digitization presents challenges due to its complex microstructure and numerous abbreviations. The goal is to systematize its organization, clarify the abbreviations, and enhance its usability. This study explores both metalexicographical and linguistic aspects to optimize the dictionary's function in digital format.

keywords: bilingual lexicography, digitization, abbreviations

## 1. Introducción

El *Nuovo dizionario spagnolo-italiano e italiano-spagnolo* de Lucio Ambruzzi (*DBA*) es una obra bilingüe, pasiva y general con elementos enciclopédicos, cuya primera edición salió a la luz en 1948 y 1949, respectivamente, primer y segundo volumen (Bermejo 2008).

El *DBA* forma parte de los materiales lexicográficos bilingües español-italiano e italiano-español sometidos al proceso de retrodigitalización (Khemakhem 2018), cuyo objetivo es ponerlos a disposición en la red. Este proceso forma parte de un proyecto que busca conformar el primer *Tesoro digital de la lexicografía bilingüe español-italiano*. El *DBA*, a diferencia de otras obras lexicográficas cronológicamente anteriores, presenta una gran complejidad microestructural, lo cual obligó al autor, en busca de unas pautas coherentes de ordenación, a diseñar un sistema tipográfico que, aplicado a las abreviaturas, se tradujo en una multiplicidad y variedad de formas. Su esfuerzo, sin duda meritorio, no alcanza, sin embargo, la sistematicidad deseada, debido a varios factores entre los que destacaremos por un lado la sobreabundancia de información, así como su amplia diversificación (enciclopédica, didáctica, cultural, curiosidades, etc.) (Calvi 2006) y, por otro lado, la estrechez que presentan los límites de las páginas de un volumen impreso.

Todo ello supone un reto a la hora de realizar el proceso de transcripción y codificación del *DBA*. Dicho desarrollo se ha articulado en distintas fases: la primera se ha concentrado en el estudio de la macroestructura y en el análisis de la aplicación de los criterios declarados por el autor en la microestructura; la segunda se ha dedicado al establecimiento de criterios de sistematización; la tercera ha consistido en la transcripción mediante la herramienta *Transkribus*, seguida de la revisión y corrección de los errores generados por el programa; y, por último, una cuarta fase de uniformización.

Para ello, centrándonos en la segunda fase, se han determinado los criterios para lograr sistematicidad y eficacia en la ordenación de la microestructura. Entre ellos, y a partir de los elementos utilizados en el *DBA*, se han compensado las faltas por omisión, repetición o incongruencia en su función, a fin de lograr la homogeneidad deseada. Tras un profundo análisis de las numerosas abreviaturas utilizadas, se ha determinado cuáles de ellas son efectivamente marcas de uso (Hausman 1989, Valero Gisbert 2016, 2018, 2020) o funcionales para la relación interna de los elementos lexicográficos, y, por tanto, se han mantenido como abreviaturas. En cambio, las demás se han explicitado logrando de este modo

superar la dificultad de la lectura y comprensión del artículo lexicográfico, favoreciendo, con ello, su interpretación cabal y su eficacia a la hora de seleccionar los equivalentes. Lo que aquí se presenta corresponde al resultado de la transcripción de la letra A en ambos volúmenes del diccionario, por lo que las decisiones adoptadas se reflejan específicamente en la transcripción de dicha letra.

En definitiva, en este artículo abordamos las principales cuestiones metalexicográficas, lingüísticas y ecdóticas que caracterizan la problemática específica de la digitalización del diccionario de L. Ambruzzi.

## 2. El diccionario

Entre los principales actores de la lexicografía bilingüe, destaca la figura de Lucio Ambruzzi. Como apunta Bermejo (2008: 201), estamos ante “un gran estudioso, lexicógrafo, gramático y filólogo, cuyos campos de interés se centraban en la lengua y se extendían a la literatura y a la cultura en general.” Esta obra ha sido objeto de un estudio pormenorizado por parte de Bermejo (2008), quien realizó un detallado análisis desde su génesis, pasando por los aspectos tanto macroestructurales como microestructurales que la definen.

Brevemente y en relación con nuestro estudio, recordamos algunos elementos especialmente interesantes para afrontar, en este estadio de la investigación, la transcripción de la letra A, dada la complejidad y la abundancia de datos.

La idea de crear esta obra no fue una iniciativa del propio autor, sino de la editorial que en distintas ocasiones lo invitó a realizarla. El diccionario se divide en dos volúmenes y si bien inicialmente aparece con el título referido a cada lemario, posteriormente, en el año 1952, se unificó en “*Nuovo dizionario Spagnolo-Italiano e Italiano-Spagnolo*” (Bermejo 2008: 132).

Esta partición se debe al diferente receptor a quien se destinaba, ya que según Bermejo (2008: 151) “la división en dos volúmenes con los aparatos críticos escritos en lenguas diferentes, al igual que las marcas, demuestra la intención de dirigirse a usuarios de dos lenguas distintas.”

En cuanto a las fuentes del diccionario, Bermejo explica que el autor declara en la introducción que se basó en obras de prestigio de distintas editoriales. Sin embargo, esta circunstancia no significa que no se pueda considerar como una obra de autor, pues como apunta la estudiosa (2008: 130), “fue él, y no la editorial, quien decidió cuántos y qué lemas incluir [...]”.

Esta breve introducción a la obra y su autor nos permite contextualizar su

génesis y comprender con mayor profundidad el proyecto y las intenciones que lo sustentan en relación con las problemáticas planteadas. A continuación, exponemos las dificultades encontradas durante el proceso de transcripción, así como los criterios adoptados para garantizar una mayor claridad.

### 3. Abreviaturas

El diccionario cuenta con dos listados de abreviaturas, uno para cada volumen, que preceden al lemario.

La sección 'Abbreviature' del volumen italiano-español registra 438; mientras que la del volumen español-italiano recoge un número menor, en concreto 337.

El listado no es circular, no son siempre las mismas, ni coinciden en la forma. En algunas ocasiones, encontramos para el mismo significado dos abreviaturas, por ejemplo, en el I volumen (sección esp-ita) Ambruzzi presenta para 'accusativo' las formas: 'acc.' y 'accusat.', mientras que en el II volumen (sección ita-esp) podemos leer, para referirse a la albañilería, las formas 'Alb.' y 'Albañ.'. Son marcas de distinto tipo, a veces se refieren a lenguajes de especialidad, o son gramaticales, o geográficas, en definitiva, con diferentes funciones.

Por otro lado, en la microestructura se observa el uso de abreviaturas que no están presentes en el listado de la macroestructura:

<i>Banc.</i> (anticipazione)	1
<i>Curt.</i> (abbancare)	1
<i>Meteor.</i> (annuvolamento)	1
<i>Telef.</i> (autocommutatore)	1
<i>Bols.</i> (arbitraggio)	1
<i>Cul.</i> (animella)	1

Fig. 1. Abreviaturas con el índice de frecuencia pertenecientes al volumen italiano-español (1949)

<i>Mur.</i> (abanto – alzar)	2
<i>Fis.</i>	18
<i>Farm.</i>	18
<i>Sport.</i> (andino – arquero)	2

masch. (algún)	1
<i>patron.</i> (Álvaro)	1

Fig. 2. Abreviaturas con el índice de frecuencia pertenecientes al volumen español-italiano (1948)

Las implicaciones que se derivan de la numerosidad de abreviaturas, junto con otras no presentes en la macroestructura e incluso el uso de diferentes formas para un mismo fenómeno, representan un aspecto complejo en la transcripción mediante el *software Transkribus*. Sirva como ejemplo la diversidad de formas para referirse a la fraseología que, como explica Bermejo (2008: 161), se encuentran: “[...] *fr.* (frase, locuzione), *loc.* (locuzione), *mod.* (modismo, modo di dire) [...].”

Esta problemática nos ha llevado a plantearnos la necesidad de establecer una serie de criterios orientados a sistematizar dichos elementos para lograr una mayor claridad microestructural en la presentación de la información. Aunque es cierto, en palabras de Bermejo (2008: 158), que desde el punto de vista tipográfico, Ambruzzi “realiza un gran esfuerzo en crear y mostrar la diversificación de los distintos tipos de información que da en el artículo lexicográfico a través de características tipográficas y de sus posibilidades combinatorias”, este cuidado por diferenciar la información contenida en el lemario se ha traducido en realidad en una dificultad a la hora de mantener la sistematicidad necesaria en toda la obra y consecuentemente en la transcripción.

Por lo que se refiere a las marcas diatópicas, la profusión de indicaciones representa, sin duda, una riqueza, pero a su vez supone un reto en términos de sistematización, especialmente si se consideran los recursos de los que el autor dispuso en su época. Para una visión más profunda, remitimos a Valero Gisbert (2020: 169-170).

#### 4. Criterios adoptados para la transcripción del diccionario de Lucio Ambruzzi (1948-49)

En la segunda fase, hemos determinado una serie de pautas para lograr un orden coherente y uniforme de las abreviaturas.

Como criterio general y por lo que respecta a las abreviaturas presentes en la lista de Ambruzzi, si no presentan ambigüedades, se mantienen con la misma forma. Hemos constatado que, en ocasiones, utiliza indistintamente la abreviatura en español o italiano, por ejemplo, para la expresión ‘etcétera’ alterna ‘etc.’ y ‘ecc’.

La elección de una lengua u otra parece que se debe al contexto en que la usa. Es el caso de los ejemplos en italiano para los que el autor opta por la abreviatura en italiano y viceversa. No se ha intervenido en estos casos, ya que el dato puede ser de interés para investigaciones futuras.

Hemos decidido registrar algunas abreviaturas en su forma completa para una mayor claridad en los dos volúmenes del diccionario. Sospechamos que Ambruzzi las dejó abreviadas por economía espacial. Es el caso de Acad. → Academia; antig. → antiguo o antiguamente; confeder. comerc. → confederación comercial; Conjúg. c./Conjúg. como/ conj. como → conjúgase como; conson. → consonante (lemma a); corresp. → corresponde; determ. → determinados; dipt. → diptongo; n. de 7 emper. rom. ...incl. → nombre de 7 emperador romano ...incluido; esdrúj. → esdrújula; especialm. → especialmente, entre otras. En el volumen español-italiano contamos 97, mientras que son un total de 151 abreviaturas en el italiano-español. A modo de ejemplo, la forma abreviada 'pal.' se ha transscrito en su forma plena: 'palabra'.

Por lo que se refiere a abreviaturas que se encuentran listadas pero que no usa en la microestructura, como por ejemplo 'patr.' y 'patron.', se ha optado por no intervenir y dejar que en futuros estudios se desambigüe el significado.

Para una mayor claridad, hemos seguido este mismo criterio con otras que, en cambio, sí estaban en la lista de Ambruzzi. En la microestructura hemos contabilizado un total de 9 en la sección ita-esp, se puede observar bajo el lema "arginale [...] adj. ref. al dique.", mientras que son mucho más numerosas en la sección esp-ita, pues se cuentan 31 abreviaturas.

#### 4.1 *Abreviaturas con errores en la microestructura*

Se trata de formas incluidas en la lista de Ambruzzi que contienen errores en su utilización dentro de la microestructura. En estos casos el criterio adoptado ha sido el de la corrección, como se puede observar en el lema "avuto *pd* habido, tenido, poseído, recibido" donde ha sido corregido con la forma 'pp.'. Por otro lado, anotamos el empleo de dos formas para señalar sinónimos, la abreviatura más usada es 'Sin.' con mayúscula o minúscula, en cambio Ambruzzi escribe 'Si-non' y 'sinon'. En este caso se ha uniformado con la forma 'Sin.'.

#### 4.2 *Abreviaturas referidas a la marcación*

Se observa que en numerosas ocasiones la abreviatura aparece entre paréntesis, otras en cursiva o redonda sin una aparente explicación. No se ha modificado, ya que puede ser importante para futuros trabajos de investigación.

A propósito de las marcas diafrecuentes, se ha intervenido, principalmente, en las que resultaban poco comprensibles como ‘p. us.’, ya que, en el primer volumen, se refiere a ‘più usato’, mientras que en el segundo se refiere a ‘poco usado’. Por lo tanto, se han sustituido con ‘più us.’ (esp-ita) y ‘poco us.’ (ita-esp). La forma ‘men. us.’ se ha transcrita con ‘menos usado’ así como ‘n. us.’ por ‘nombre us.’, del mismo modo ‘más común’ se ha uniformado con ‘más us.’

En lo que respecta a las marcas cronológicas, la abreviatura ‘ant.’ presenta cierta ambigüedad, ya que puede interpretarse como ‘antiguo’, ‘antiguo’ o ‘antillano’. En este caso se ha optado por conservarla en su forma original, considerando que futuros estudios sobre este aspecto podrían arrojar datos de interés.

Se observa el caso de la marca gramatical correspondiente a ‘impersonal’ que en la lista aparece como ‘impers.’, mientras que Ambruzzi emplea en algunas ocasiones la forma abreviada ‘imp.’. En estos casos hemos cambiado esta última uniformándola a la marca que está en la lista, es decir, ‘impers.’ También la utilizamos para aludir a ‘imperio’, en este caso la hemos desarrollado eliminando la ambigüedad presente en ‘imp.’.

Por lo que se refiere a las marcas diaevaluativas o pragmáticas se han mantenido tal como las emplea el autor.

En el volumen español-italiano, observamos que la marca diactéctica correspondiente al ámbito de la medicina presenta formas diferentes: se presenta de una manera en la lista de abreviaturas, y adopta una forma diferente en el cuerpo del diccionario. En la lista emplea ‘med.’ para referirse al ámbito médico, mientras que en la microestructura aparece ‘medic.’ En estos casos se ha optado por registrar la marca completa.

En el mismo volumen, las marcas diatópicas se han simplificado y uniformado de la manera siguiente: ‘spagn.’ → ‘sp.’; ‘ital.’ → ‘it.’

En el caso de la abreviatura ‘rom.’, se ha optado por conservar íntegramente las formas utilizadas por Ambruzzi, con la única excepción de esta marca, interpretada como ‘romano’, decisión tomada siempre en aras de una mayor claridad. Un ejemplo representativo se encuentra en la entrada “*Apuleio* Lit. Lucio Apuleyo, escritor rom.”

Por último, hemos encontrado alguna marca abreviada (ausente en la lista)

que hemos decidido mantener precisamente para no contaminar posibles estudios futuros, como la referida al dólar bajo el lema “*aquila [...] moneda de a 10 doll. de Estados Unidos [...]*”.

Para terminar, cabe señalar que se han eliminado restos de frases que probablemente por distracción se habían quedado en el texto.

## 5. Herramientas utilizadas para la transcripción mediante Transkribus

El proceso de retrodigitalización de una obra implica varias etapas, a veces largas y complejas, que son posibles gracias a las herramientas desarrolladas en el ámbito de las Humanidades Digitales (HD).

Aplicando la distinción de Bazzaco (2020), la primera fase es la digitalización, es decir, el escaneo de la obra en papel en imágenes procesables por un ordenador. A continuación, sigue la digitización: “la transformación de un texto analógico en un texto electrónico en formato mrf (*machine readable form*)” (Bazzaco 2020: 536).

Esta segunda etapa se simplifica gracias a los sistemas OCR (*Optical Character Recognition*) y HTR (*Handwritten Text Recognition*) que permiten el reconocimiento automático y semi-automático de las imágenes del texto y su transcripción a texto electrónico, sustituyendo en parte la transcripción manual (Bazzaco 2018). Estas herramientas, ya útiles desde las primeras fases de desarrollo, en las últimas décadas han mejorado cada vez más gracias a la integración de algoritmos de Inteligencia Artificial (IA) y aprendizaje automático (*machine learning*) que permiten al *software* aprender y mejorar a partir de los datos que proporciona el usuario (Bazzaco 2020), consiguiendo así márgenes de error cada vez menores.

Entre los diferentes recursos disponibles, se ha decidido utilizar la plataforma Transkribus<sup>1</sup>, un *software* en continua actualización que, gracias al uso de sistemas HTR e IA, reconoce tanto textos manuscritos como impresos y permite entrenar modelos personalizados que reconocen las peculiaridades del texto propio. La plataforma nace en 2015 en el seno del proyecto READ y ahora está dirigida por la cooperativa READ-COOP SCE cuyo objetivo es: “*to provide a comprehensive range of tools and services that empower researchers, institutions, and individuals to collaboratively discover and explore the rich tapestry of history*”<sup>2</sup>. Dicha cooperativa cuenta con más de 200 miembros en 30 países del mundo y más de 300 000

1 Cfr. <https://www.transkribus.org/it>

2 Cfr. <https://shorturl.at/6lmiv>

usuarios, lo que lo convierte en uno de los programas de transcripción automática más utilizados.

Para la transcripción de la letra A de los dos volúmenes de la primera edición del diccionario de Ambruzzi, se ha seguido un proceso metodológico similar al que se ha seguido en otros proyectos previos (e.j. Bazzaco 2018, 2020, Nalesto 2024). Tras el escaneo de los dos volúmenes, se procedió a trabajar en la plataforma *Transkribus*, comenzando con la carga de las imágenes. El siguiente paso fue el *Layout Analysis*, es decir, la segmentación de cada página en varias áreas para que el *software* leyera el texto en el orden apropiado. En el caso de nuestra obra, como es típico en los diccionarios bilingües, la página está dividida en dos columnas, por lo que fue necesario segmentarla en dos regiones para que se leyera todo correctamente. A continuación, habría sido necesario transcribir manualmente unas páginas para crear el *ground truth*, la base para el desarrollo de un modelo personalizado. Sin embargo, como sugiere la misma plataforma de *Transkribus*, las páginas de *ground truth* se transcribieron automáticamente con un modelo público disponible para todos los usuarios. De hecho, tanto para el primer volumen (esp-ita) como para el segundo (ita-esp), se utilizó inicialmente el modelo *TRANSKRIBUS PRINT M1* entrenado con 10 281 páginas y 14 idiomas diferentes, español e italiano incluidos. Las páginas transcritas con este modelo se corrigieron manualmente para lograr transcripciones perfectas y luego se utilizaron como entrenamiento para los dos primeros modelos personalizados.

A pesar de los diferentes problemas y procedimientos, el resultado final ha sido una transcripción correcta de la primera letra del diccionario, descargada de la plataforma en formato TXT.

A continuación, se analiza más detalladamente el proceso de transcripción y entrenamiento, los errores que surgieron en ese proceso y las correcciones que se hicieron manualmente para lograr una transcripción correcta de la letra A.

## 6. Primer volumen SPA-ITA

Este primer volumen de la primera edición del DBA se titula *Nuovo Dizionario Spagnolo-Italiano* y, aunque por cuestiones de *copyright* siempre aparece como publicado junto con el segundo volumen en 1949, en realidad apareció un año antes (Bermejo 2008).

En cuanto a la macroestructura, cuyas páginas están marcadas con números romanos, hay 4 partes: la 'Presentazione', la 'Pronunzia dello spagnolo' y las 'Abbreviature' están escritas en italiano mientras que la 'Advertencia' está en español porque el autor se dirige a los hispanohablantes explicando el sistema de la acentuación italiana y cómo les ayudará a entender la pronunciación de las palabras.

Por lo que respecta a la microestructura, cuenta con un total de 1 095 páginas y aproximadamente 65 000 entradas. Como ya se ha explicado, dada la extensión y la abundancia de información, para esta primera parte del trabajo nos hemos centrado únicamente en la digitación de la letra A, que consta de 147 páginas y 9 987 entradas.

### 6.1 *Transkribus: problemas y resultados*

A continuación, mostramos la metodología empleada para la transcripción. En primer lugar, se escaneó el diccionario, seguidamente se subieron las imágenes a *Transkribus* para empezar el proceso de transcripción automática. La parte transcrita con el modelo público *TRANSKRIBUS PRINT M1* fueron las primeras 14 páginas, es decir, el 10% del total. En el caso del primer volumen, el modelo entrenado con estas primeras páginas fue tan eficaz que no se consideró necesario entrenar nuevamente el *software*. Seguidamente, se presentan los principales errores antes y después del entrenamiento para mostrar su eficacia.

La transcripción con el modelo público contaba 1 358 errores, con un mínimo de 60 fallos en la página que contaba un número menor y un máximo de 133 equivocaciones en la página con mayor cantidad.

En cuanto a la tipología de erratas pueden identificarse 5 categorías diferentes:

#### 1. Ausencia de signos de puntuación

En todas las páginas faltan casi todas las barras verticales (|) que el autor utiliza para dividir las diferentes acepciones de un lema, y las barras dobles (||) que se emplean cuando una palabra cambia también de categoría gramatical (Bermejo 2008). De hecho, del total de los errores en estas 14 páginas, 583 son omisiones de barra vertical.

Además, a menudo omite los guiones para dividir palabras al final de las líneas y rayas (—) utilizadas como sustitutos del lema en los ejemplos. Por lo que se refiere a puntos, comas, paréntesis y signos de exclamación e interrogativos, normalmente el programa los transcribe.

## 2. Equivocación entre signos de puntuación

Como hemos adelantado, el guion (salvo cuando se omite) se transcribe con el símbolo ‘-’. Este es el error más común de este grupo y se repite en todas las páginas con un total de 155 veces. Además de confundir la coma con el punto en 11 páginas, solo se equivoca en otras ocasiones alterando signos parecidos como los exclamativos de apertura o cierre con barras o doble barras verticales o puntos y comas y dos puntos con signos de exclamación.

## 3. Ausencia de letras o abreviaturas

En esta primera transcripción sin entrenamiento, el programa no leía algunas letras de ciertas palabras, como por ejemplo la ‘i’, la ‘u’ o la ‘o’. Por otro lado, tampoco ha reconocido en la transcripción algunas marcas gramaticales, como la ‘r.’ para los verbos reflexivos, ‘s.’ para los sustantivos, ‘agg.’ para marcar los adjetivos y en cuatro casos no aparece la abreviatura ‘fig.’ que señala el sentido figurado de las palabras.

## 4. Equivocación entre letras

Este es el grupo donde aparece la mayor variedad de errores, aunque no se repiten con frecuencia. Salvo en los casos de confusión entre ‘l’ y ‘f’ en la marca gramatical ‘sf.’, así como entre las letras ‘l’ e ‘i’—errores que se registran con una frecuencia de 41 y 31 ocurrencias respectivamente—, el programa presenta numerosos errores en la transcripción de letras, sustituyéndolas por otras. No obstante, dichos errores no se repiten de manera sistemática.

En tres ocasiones acorta algunas abreviaturas escribiendo solo ‘f.’ o ‘g.’ en lugar de ‘fig.’, o ‘a.’ para ‘agg.’ y dos veces escribe la letra ‘j’ confundiéndola con los signos de exclamación de apertura y cierre. Altera algunas letras mayúsculas, como la ‘T’ en lugar de la ‘T’ o la ‘T’ en lugar de la ‘i’; sin embargo, en general, presenta dificultades en el reconocimiento de determinadas letras, que a menudo confunden entre sí. Por ejemplo:

- ‘t’ en lugar de ‘i, í, f o l’
- ‘é’ en lugar de ‘v, c, o’
- ‘o’ en lugar de ‘g, e, o, c’
- ‘m’ escrita como ‘n, ni, nt’

## 5. Desaciertos en la acentuación

Siendo un diccionario español-italiano donde el autor marca también la sílaba

tónica de algunas palabras italianas, hay muchísimos acentos que el *software* no siempre transcribe o reconoce bien. El error más común es la tilde en la letra ‘í’; este aparece, por lo menos una vez en todas las páginas, con un total de 57 veces.

Todos los errores en estas páginas se corrigieron manualmente y luego constituyeron la base para entrenar un nuevo modelo capaz de leer mejor las letras y los símbolos de este volumen.

En las 133 páginas siguientes se utilizó el nuevo modelo personalizado y de 1 358 errores en 14 páginas se obtuvo un resultado de 1 860 errores en 133. Por lo tanto, proporcionalmente, el sistema entrenado funcionó mejor.

En cuanto a la tipología, los errores son similares a los de antes del entrenamiento, pero el número ha disminuido notablemente. El más común sigue siendo la ausencia de barra vertical ‘|’, que ahora aparece en 128 páginas con un total de 475 ausencias. Si antes el segundo tipo de equivocación era la omisión de la doble barra, ahora son las rayas (101 veces) y los guiones para dividir la sílaba al final de las líneas que ya no se confunden con el símbolo ‘—’. Sin embargo, se omiten en 61 páginas con un total de 105 veces. Las comas y los puntos resultan ausentes 42 veces y hay muy pocas omisiones de paréntesis, signos de exclamación, puntos y comas o dos puntos.

Por lo que respecta a los errores de confusión entre letras y signos de puntuación, se presentan también después del entrenamiento. En cuanto a la equivocación entre signos, como hay más páginas, emerge una mayor variedad de errores. El más repetido es la barra vertical en lugar de la doble barra, que aparece 25 veces, seguido de la confusión entre coma y punto. La barra vertical se transcribe algunas veces también en lugar de punto, raya, guion, coma, paréntesis o las letras ‘f’ o ‘í’, así como la coma confundida con paréntesis, guion, punto y coma o dos puntos. En las páginas de *ground truth* no aparecen comillas angulares («»), a veces utilizadas por el autor para las citas, y por eso, cuando se dan, la plataforma las transcribe de forma distinta: una vez con paréntesis ‘(’, y dos veces con ‘s’ y ‘v’; lo mismo con los corchetes ([ ]), que dos veces se confunden con paréntesis de apertura y otras dos con la letra ‘l’, y con los números que se mezclan entre ellos. El número ‘4’ se transcribe una vez con la letra ‘l’, el ‘58’ con ‘ss’, el ‘7’ con ‘T’ y el número ordinal ‘2’ que aparece una vez como ‘?’ y otra como ‘ze’. En cuanto a las letras, se confunde con otras similares como la ‘f’ en lugar de ‘l’ o la ‘l’ en lugar de ‘í’, los dos errores más comunes, mientras que otras letras aparecen ocasionalmente por error. Como antes, casi todas las erratas con los acentos son la ausencia de tilde o confusión entre acento agudo y grave, de hecho, la más común sigue siendo la omisión de tilde sobre la ‘í’. Sin embargo, en estas páginas

se encuentran dos acentos que no estaban en las primeras catorce: la diéresis en la ‘ú’ y el circunflejo en la ‘í’. El primero casi siempre se transcribe como ‘ú’ o ‘ú’ en lugar de ‘ú’, mientras que ‘í’ aparece algunas veces sin acento.

El único tipo de error que no aparece en la transcripción con el modelo público, y en cambio sí se encuentra después, es la inclusión de signos de puntuación y letras donde, en realidad, no las hay. A lo largo de las 133 páginas, el programa añade dos ‘í’ y una ‘c’, así como 29 barras verticales, 25 guiones, 15 rayas, un paréntesis y algunos puntos donde el original no los presenta.

En general, gracias a este breve resumen del número y el tipo de errores que produce *Transkribus* antes y después del entrenamiento con la IA, se puede ver cómo funciona este último y conduce a una disminución significativa de los errores, haciendo que el proceso de transcripción automática sea eficaz y útil durante la digitalización.

## 6.2 *Detección y corrección de erratas en la versión en papel*

En el apartado anterior hemos hablado de los errores cometidos por el programa de transcripción automática. Ahora, en cambio, se presentan algunos desaciertos e incongruencias detectados en el volumen impreso, que se corrigieron y modificaron manteniendo el criterio de claridad y uniformidad. Durante la corrección de la transcripción se observó que la obra presenta erratas, a veces probablemente de impresión, y contradicciones con lo declarado en la macroestructura.

Entre los principales cambios realizados se encuentran los referidos a las abreviaturas y marcas de uso. En las entradas del diccionario se observa un considerable número de palabras que, si bien no corresponden a marcas de uso, se abreviaron seguramente por razones de espacio y formato de impresión. Sin embargo, dado que una edición digital no tiene tales limitaciones, se optó por desarrollar la mayoría de estas formas abreviadas con el objetivo de facilitar la comprensión del texto. Muchas de las palabras que se han extendido no figuran en la lista de ‘Abbreviature’ proporcionada por el autor. Excluyendo las repeticiones, se han identificado 97 términos abreviados que no están registrados en dicha lista. Algunos ejemplos son: ‘alfab.’ por ‘alfabeto’, ‘condiz.’ por ‘condizione’, ‘esclam.’ o ‘escl.’ por ‘esclamazione’, y ‘mis.’ por ‘misura’, esta última con tres apariciones. Asimismo, se hallaron abreviaturas que sí están incluidas en el listado original, así como formas abreviadas que difieren de las registradas por el autor. En ambos casos, dado que no se trataba de marcas de uso, se ha decidido igualmente

expandir las palabras. En el primer caso—abreviaturas coincidentes con la lista—se contabilizaron 164 casos (incluyendo repeticiones), entre los cuales destacan: ‘dic.’ por ‘dicesi’ (la más frecuente, con 50 apariciones), ‘qc.’ y ‘qu.’ por ‘qualcosa’ y ‘qualcuno’, ‘c.’ y ‘ab.’ por ‘città’ y ‘abitanti’, o ‘sim.’ por ‘simile’. En el segundo caso—abreviaturas con formas diferentes a las del listado— se registraron 31 casos. Algunos ejemplos representativos son:

- ‘accusat.’ para ‘accusativo’ que es en la lista aparece como como ‘acc.’ y ‘accusat.’
- ‘medic.’ para ‘medicinale’ que en la lista se abrevia ‘med.’
- ‘abit.’ para ‘abitante’ que, como acabamos de ver, en la lista aparece como ‘ab.’
- ‘tradiz.’ se abrevia ‘trad.’ para ‘tradizione’

En cuanto a las marcas de uso, casi todas están en el listado de Ambruzzi, sin embargo, figuran numerosas incoherencias en la forma. Como ya se ha explicado anteriormente, el sistema de abreviaturas creado por el autor es muy amplio y, ya en la macroestructura, presenta distintas formas para una misma información. Por otro lado, la situación se complica aún más por el hecho de que en la microestructura, además de las numerosas formas de la lista, se encuentran otras adicionales. A continuación, se presentan los tres escenarios encontrados y los cambios que se han realizado:

1. En ‘Abbreviature’ aparecen dos formas, pero en la microestructura se utiliza casi siempre solo una. En este caso se vio cuál era la marca más frecuente y se uniformó todo según esta. Aunque la tabla se refiere únicamente a la letra A, las decisiones se han tomado sobre la base de la frecuencia. Por ello, si bien la marca ‘gal.’ aparece solo cinco veces en la letra A, es en realidad la más frecuente a lo largo de todo el diccionario.

LISTA				
<i>ital. – it.</i>	it.	16	ital.	5
<i>spagn. – sp.</i>	sp.	16	spagn.	7
<i>gall. – gal.</i>	gal.	5	gall.	24
<i>Ven. – Venez.</i>	Venez.	70	Ven.	1
<i>Comm. – Com.</i>	Com.	41	Comm.	11
<i>Eq. – Equat.</i>	Equat.	69	Eq.	4
<i>Ed. – Edil.</i>	Edil.	7	Ed.	1

Fig. 3 Marcas presentes en la lista con dos formas

2. Marcas de uso que en la lista tienen una forma y en la microestructura otra, incluso, en raras ocasiones, se usa otra diferente que no está presente.

LISTA	ENTRADAS	
Arg.	Argent.	1
am.; amer.	americ.	2
dif.	difett.	2
Fil.	Filipp.	2
P.p.; p.p	pp.	11
Ven.	Venat.	1
Par.	Parag.	1
lat.	l.	1
superl.	superlat.	1
	sup.	8
St.	Stor.	3
loc.	locuz.	2
	l.	1
Fot.	fotogr.	1
et.; etim.	etimol.	1
inv.	invar.	1
Gram.	Gramm.	4
acc.; accr.	accresc.	2
Am. M.	Amer. Merid.	5
per est.	per estens.	2
f.; fem.	femm.	5
Mess.	Mes.	1
Dom.; Domen.	Domin.	1
antill.	antil.	2
Ant.	Antil.	5
spr.	sprech.	1
con.; coniug.	si con.	3
Filos.	Fil.	1
Metal.	Met.	1
fr. prov.	fr. proverb.	1

Fig. 4 Marcas diferentes en las entradas

Aquí, son escasas las ocasiones en las que aparece la forma que no está en la lista, en estos casos se cambió por la de la lista que es también la más común.

3. Marcas que en el listado tienen una forma que apenas se utiliza porque es más frecuente otra. En estos casos se ha dejado la marca más utilizada, aunque no figure en la lista.

Estas son:

- ‘Met.’, marca diatécnica para ‘Metereología’ que solo aparece 5 veces en todo el volumen, sustituida por la más usada ‘Meteor.’
- ‘C. Am.’, marca diatópica para ‘America Centrale’, se usa solamente 26 veces en todo el diccionario frente a las 368 ocasiones en las que aparece ‘C. Amer.’
- ‘Per.’, marca diatópica para ‘Perú’ aparece solo una vez, mientras que las otras veces se escribe por entero.

Las únicas dos excepciones que no entran en estas categorías son la abreviatura ‘Centroam.’ que no se cambia por la marca diatópica presente en la lista, sino con la marca más frecuente, es decir, ‘C. Amer.’, y las dos veces que figura ‘Cub.’, ausente en la lista, ya que como marca diatópica para referirse a Cuba se utiliza el nombre entero.

Además de los cambios con las abreviaturas y marcas de uso, también se corrieron otros errores encontrados en la primera letra.

### 1. Fallos de puntuación

En las entradas, el autor utiliza diferentes signos de puntuación que, a veces, son incorrectos. Muchas veces faltan ciertos signos como el punto al final de una entrada o después de una abreviatura, la coma para separar los equivalentes, el guion para dividir una palabra al final de la línea o los paréntesis de apertura o cierre.

En ese grupo se pueden incluir también la ausencia del guion para marcar la desinencia del femenino.

### 2. Errores de acentuación

Con respecto a las palabras españolas, hay tantos lemas como palabras dentro de las entradas que llevan tilde donde no deberían como: ‘afluír’ → ‘afluir’, ‘Albanía’ → ‘Albania’ o ‘algortimía’ → ‘algoritmia’, y otras que no la llevan (e.j. ‘actínico’ → ‘actínico’ o ‘aligator’ → ‘aligátor’). Por otra parte, en el diccionario de Ambruzzi, además de los acentos españoles, el autor marca también la sílaba tónica de algunas palabras italianas para ayudar a quienes no conocen bien el idioma a entender la pronunciación correcta. En la ‘Advertencia’ afirma que pon-

drá el acento agudo sobre las vocales cerradas (é, í, ó, ú) y el grave sobre las vocales abiertas (à, è, ò). Sin embargo, se presentan algunas excepciones en las que no utiliza el acento correcto con respecto a lo que declara en la macroestructura. Por ejemplo: 'rullío' → 'rullío', 'fúcile' → 'fucíle' o 'cùmulo' → 'cúmulo'. En total se cuentan 7 fallos en los acentos italianos.

### 3. Marcas y palabras en español

En el primer volumen esp-ita, los lemas están en español y las marcas y las entradas en italiano con excepciones de los ejemplos en español. A pesar de eso, en algunos casos se encontraron palabras escritas en español donde deberían haber estado en italiano. Ocurre dos veces con la palabra 'Córdoba' y una vez con 'Castilla' que deberían ser 'Cordova' y 'Castiglia' respectivamente, y con las marcas:

- 'Chile' para 'Cile' (8 veces)
- 'Cile y Esp.' para 'Cile e Sp.'
- 'desp.' (despectivo) para 'dispr.' (dispregiativo)

El error más significativo que no se ha modificado es la presencia de la entrada "ascetismo" escrita toda en español.

### 4. Errores de imprenta

Se incluyen aquí todos los casos en los que una palabra o frase presenta errores ortográficos debidos a la omisión de una letra—como en 'éssicare' por 'essicare' u 'olfato' por 'olfatto'—o a la sustitución de una letra por otra, como en 'doce' por 'dove', 'alealoideo' por 'alcaloideo' o 'agalogia' por 'agatologia'. Este tipo de errores se observa ocasionalmente también en las marcas, por ejemplo: 'sgg' en lugar de 'agg', o 'Zoot' en lugar de *Zool.*

Además, se señala que en el primer volumen se encontró un lema repetido dos veces y un orden alfabético incorrecto que, aunque en un tesoro lexicográfico no resulta un problema, en la corrección de la transcripción se ha corregido.

A pesar de los varios cambios que se hicieron para que el volumen resulte lo más uniforme y claro posible, hay algunos aspectos que no se han abordado en este trabajo, como, por ejemplo, los significados de la marca 'ant.' o el aspecto metalexicográfico relacionado con el uso de la puntuación, las mayúsculas, la cursiva y la negrita, que presentan varias incongruencias.

## 7. Segundo volumen ITA-SPA (1949)

El segundo volumen de la primera edición del DBA se titula *Nuovo Dizionario*

*Italiano-Spagnolo* publicado en el año 1949 (Bermejo 2008).

El volumen está compuesto por 1 337 páginas que comprenden una introducción en la que el autor presenta la historia de la lengua italiana con anotaciones sobre la fonética, la morfología y la sintaxis.

Además, dedica un apartado a la pronunciación de vocales y consonantes, una sección sobre las reglas de acentuación, subrayando también que la lengua italiana no tiene reglas precisas de acentuación gráfica como el español. A continuación, incluye otras notas sobre el uso de preposiciones, partículas relativas y nexos, así como sobre las fuentes utilizadas para elaborar el volumen y, por último, la lista de abreviaturas. Al final del volumen, añade un anexo con algunas erratas. Como ya se ha explicado, dada la extensión y la abundancia de información, en esta primera parte del trabajo —al igual que en el primer volumen— nos hemos centrado únicamente en la digitación de la letra A, que consta de 131 páginas y entradas.

## 7.1 Transkribus: *problemas y resultados*

Al igual que para la transcripción del primer volumen, para la del segundo se ha utilizado el programa *Transkribus*.

Inicialmente, se utilizó el modelo del programa, *TRANSKRIBUS PRINT M1*, a partir del cual se detectaron numerosos errores.

Los principales problemas se debieron principalmente a la escasa legibilidad del PDF, que no permitía distinguir letras aparentemente similares; a los signos añadidos posteriormente a lápiz o bolígrafo; a las líneas discontinuas especialmente en correspondencia del guion largo, utilizado para sustituir el lema italiano en los ejemplos, seguido de la pleca; y a la ausencia o inversión de acentos tanto en palabras italianas como españolas.

A diferencia del primer volumen, la transcripción del segundo exigió la creación de tres modelos más en los que, aunque mínimas, se observaron mejoras. Si se calcula el promedio de errores, los resultados obtenidos son los siguientes: promedio de página 1 a 13 (modelo *Transkribus*): 115 erratas; de página 14 a 30 (modelo I): 62 equívocos; de página 31 (modelo II): 88 faltas; de página 100 (modelo III): 55 fallos.

A continuación, la lista de los errores más frecuentes que se han corregido manualmente a lo largo de la letra A.

- —
- - (al final de renglón)
- ) → j
- / → f
- : → << >>
- ¬ → -
- 1 → i
- A → À
- a → á; à; e; g; o; c
- à → á; u
- á → à
- \* → a
- b → h
- c → e; v
- D → B
- é → á
- è → è
- e → o; ú
- f → i; t
- fg. → fig.
- í → á; f
- i → f; í; j; ú;
- í → i; s
- ifl → ¼
- il → XII
- j → f; p; h
- l → !; 3º; f; h; í; i; ñ; t; T
- L → í
- l 9y → 107
- l → l
- ll → II
- m → n
- ñ → f
- n → u
- ó → è
- o → e
- P → V

q → g  
 r → m; v  
 sfs → 678  
 t → f; i; í; l  
 u → g; ú; a; n  
 ú → u  
 v → n  
 V → y

## 7.2 Detección y corrección de erratas en la versión en papel

Además de los errores derivados de la transcripción, se detectaron numerosas erratas. Concretamente, tras la transcripción, se exportó el archivo en TXT sobre el que se realizó una revisión minuciosa para corregir lemas, abreviaciones, marcas de uso, origen, equivalentes, ejemplos y fraseología. En vista del gran número de fallos encontrados en el segundo volumen (387 totales), se optó por dividirlos en 23 grupos y establecer una tipología.

Con el fin de ser más precisos, a continuación, se presentan los grupos y la tipología acompañados por el número total de las erratas encontradas y los ejemplos más explicativos. En el caso de los grupos más numerosos se ha indicado un ejemplo por tipo.

1. División silábica incorrecta al final de renglón (9 veces): (lema ‘adito’) recur- -rido → recu- -rrido; (lema ‘anello’) epi- -scopal → epis- -copal.
2. Inversión de letras presentes en lemas, marcas, equivalentes, plurales, etc. (8 veces): (lema ‘A.E.I.O.U.’) Absburgo → Asburgo; (lema ‘affannare’) hijo → hijo; (lema ‘ammodo’) ocn → con; (lema ‘ausònico’) -ic → -ci; (lema ‘Australia’) Georg. → Geogr.; (lema ‘assiepare’) rt. → tr.; (lema ‘astuccio’) cajtaà → cajita; (lema ‘atteggiare’) acittud → actitud.
3. Falta de orden alfabético en las entradas (5 veces): sobre los lemas: ‘appillottare’ desplazado antes de ‘appinzare’; ‘animetta’ desplazado antes de ‘animismo’; ‘allotriofagía’ desplazado antes de ‘alotropía’; ‘astensionista’ desplazado antes de ‘aster’; ‘annaffiata’ desplazado antes de ‘annaffiato’.
4. Errores que quedan por descifrar: (lema ‘año’) correte.
5. Sustitución de letras que se pronuncian igual (2 veces): (lema ‘agenzia’) mensagería → mensajería; (lema ‘avventato’) lijero → ligero.
6. Mayúsculas (2 veces): (lema ‘ascrittizio’) arqueol. → Arqueol.; (lema ‘Attico’) ático → Ático, al tratarse de un nombre propio, se ha corregido con la mayúscula.
7. Uso de letras por otras (44 veces): (lema ‘aggiustatezza’) precisión → precisión; (lema

‘Alighieri’) Danre → Dante; (lema ‘automobile’) velodidades → velocidades; (lema ‘arricciare’) francir → fruncir; (lema ‘arruffato’) anmarañado → enmarañado; (lema ‘aspettativa’) expectatica → expectativa; (lema ‘addestrare’) sorvirle → servirle; (lema ‘agitazione’) commoción → conmoción; (lema ‘ameno’) extrano → extraño; (lema ‘appiccatura’) añabidura → añadidura; (lema ‘Aristarco’) gramárico → gramático; (lema ‘appoggio’) Ginn. → Gimn.; (lema ‘ärnica’) Rot. → Bot.; (lema ‘ardore’) sn. → sm.

8. Empleo de palabras en italiano en frases españolas (8 veces): (lema ‘accréscere’) famiglia → familia; (lema ‘agitare’) notizia → noticia; (lema ‘alto’) elevato → elevado; (lema ‘aggiogàbile’) che → que; (lema ‘applicazione’) ingegnería → ingeniería; (lema ‘abbominio’) e deriv. → y deriv.; (lema ‘asfissiante’) asfissiante → asfixiante; (lema ‘assuntivo’) gentilizio → gentilicio.

9. Palabras que terminan en *-tor* en lugar de *-dor* (3 veces): (lema ‘aletta’) radiator → radiador; (lema ‘apòstolo’) predictor → predicador; (lema ‘aratro’) regulator → regulador.

10. Términos españoles con acentos italianos (8 veces): (lema ‘annunciare’) està → está; (lema ‘acinoso’) glàndula → glándula; (lema ‘Aventino’) estòmago → estómago; (lema ‘Àbila’) Àbila → Ábila; (lema ‘Angelo’). Àngel → Ángel; (lema ‘Àtropo’) Àtropos → Átropos; (lema ‘Aristòfane’) comediògrafo → comediógrafo; (lema ‘ànitra’) ànade → ánade.

11. Palabras que deben separarse (3 veces): (lema ‘accozzare’) cadauno → cada uno; (lema ‘alcònico’) sunido → su nido; (lema ‘anagífride’) leñoh. → leño h.

12. Repeticiones (5 veces): antinazionale (lemma) → antinazionale (equivalente); (lema ‘aiutante’) id. id. → id.; el lema ‘atomística’ se repite dos veces, uno con acento fónico y el otro sin acento, respetando las indicaciones de Ambruzzi, hemos decidido dejar el lema con acento fónico; repetición del lema ‘avvampamento’; repetición de la marca adj. (lema ‘antipùtrido’).

13. Presencia de acentos en palabras que no los llevan (9 veces): (lema ‘addentallato’) dentículado → denticulado; (lema ‘accerito’) enrojecido → enrojecido; (lema ‘Aladino’) medio → medio; (lema ‘ametista’) amétista → ametista; (lema ‘ametría’) arritmia → arritmia; (lema ‘animaletto’) cojíjo → cojijo; (lema ‘antirelligioso’) antirreligioso → antirreligioso; (lema ‘angelicamente’) angéicalmente → angelicamente; (lema ‘accordare’) sentíimiento → sentimiento.

14. Letras o sílabas que se omiten o añaden (31 veces): (lema ‘ammazzare’) a polar → apilar; lema ‘adustar’ → adustare; (lema ‘azzardare’) quell’uomo (ejemplo) → quell’uomo; (lema ‘alessandrinismo’) alejandrínismo; (lema ‘arma’) estar bajo las, → estar bajo las a.; (lema ‘autenticità’) autencidad → autenticidad; (lema ‘asserre’) en las conjugaciones Pre. → Pres.

15. Palabras divididas (3 veces): (lema ‘affezionare’) n os → nos; (lema ‘affocare’) pasio nes → pasiones.

16. Lema en español: ‘aerificación’ → aerificazione.

17. Letras dobles (5 veces): (lema ‘abbattimento’) postracción → postración; (lema

‘apostolato’) missión → misión; (lema ‘arcifanfano’) fanfarón → fanfarrón.

18. Letras que hay que eliminar (3 veces): (lema ‘avaccio’) prontitudo → prontitud; (lema ‘avanti’) ¡marchen, mar! → marchen, ar!

19. Acento en el diptongo *-ui* (24 veces): (lema ‘accommiatare’) fuímos → fuimos (ejemplo); (lema ‘accluso’) incluído → incluido (equivalente); (lema ‘agnosticismo’) descuido → descuido; (lema ‘altruismo’) altruísmo → altruismo; (lema ‘acuto’) buído/buída → buido/buida; (lema ‘assommare’) concluīr → concluir; (lema ‘assentire’) anuīr → anuir; (lema ‘avvilire’) ruín → ruin.

20. Ausencia de acentos en palabras españolas (31 veces): (lema ‘appello’) tambien → también; (lema ‘appaltatore’) compañía → compañía; (lema ‘attingimento’) liquidos → líquidos; (lema ‘atleta’) campeon → campeón; (lema ‘attivistico’) activistico → activístico; (lema ‘avvertire’) cuidate → cuídate; (lema ‘accio’) tu → tú; (lema ‘aria’) ventilacion → ventilación; (lema ‘alzata’) mas alto → más alto.

21. Ausencia de *-s* final (verbos, artículos, sustantivos, etc.) (6 veces): (lema ‘antenora’) traidore → traidores; (lema ‘abitino’) terciario → terciarios; (lema ‘Adamástor’) la → las; (lema ‘affinché’) pueda → puedas; (lema ‘Ampolla’) lo → los.

22. Errores de imprenta en las marcas (acentuación y varias) (96 veces): iron → irón; Quim → Quím; ár. → árabe o voz árabe; Amer. → Amér.; contracc. → contrac.; Sider. → Sid.; St → Hist; Méjco → Méj.; Geograf. → Geogr.; superlat. → superl.; infinit. → infin.; transit. → tr. - intransit. → intr.; s. → sf.; anglos. → ingl.; agg. → adj.; avv. → adv.; abr. → abrev.; antig. → ant.

23. Errores de puntuación (80 veces): ausencia de paréntesis al final del artículo lexicográfico; ausencia del punto final; presencia del punto final dentro del paréntesis; sustitución de punto con la coma y viceversa; falta del guion en los plurales y al final del renglón; falta de la coma entre equivalentes; falta el punto después de las marcas; presencia de puntos, comas, paréntesis y plecas no necesarios; sustitución de la doble pleca con el punto.

## 8. Conclusiones

Tal como se ha expuesto a lo largo de este estudio, el diccionario Ambruzzi constituye una obra valiosa por la abundancia y diversidad de datos que proporciona al lector. Sin embargo, esta misma riqueza representa un desafío en el proceso de digitalización, ya que requiere la unificación de múltiples aspectos para alcanzar una coherencia interna. Esta coherencia, sin duda, no estaba al alcance del autor debido a las limitaciones de los medios manuales con los que trabajó. En este proyecto, nuestro principal objetivo ha sido resolver dichas inconsistencias, tal como se ha detallado en esta investigación.

## Bibliografía citada

Bazzaco, Stefano (2018), “El Progetto Mambrino y las tecnologías OCR: estado de la cuestión”, *Historias Fingidas*, 6: 257–72 [03/04/2025] <<https://historiasfingidas.dlls.univr.it/article/view/89/135>>

Bazzaco, Stefano (2020), “Siglo de Oro: creación de un modelo HTR basado en libros de caballerías del siglo XVI en la plataforma Transkribus”, *JANUS*, 9: 534–61 [03/04/2025] <<https://www.janusdigital.es/descargar.htm?id=160>>

Bermejo Calleja, M. Felisa (2008), “El Nuovo dizionario spagnolo-italiano e italiano-spagnolo (1948-1949) de L. Ambruzzi”, *Textos fundamentales de la lexicografía italoespañola (1917-2007)*, ed. Félix San Vicente. Milán, Polimetrica, Vol. 1: 125-97.

Calvi, Maria Vittoria (2006), “Il componente culturale del dizionario Ambruzzi di spagnolo e italiano”, *Lessicografia bilingue e traduzione: metodi, strumenti, approcci attuali*, ed. Félix San Vicente. Monza, Polimetrica: 83-10 [03/04/2025] <<https://www.yumpu.com/it/document/read/16097139/open-access-publications-eurac>>

Hausmann, Franz Josef (1989), “Die Markierung in allgemeinen einsprachigen Wörterbuch: eine Übersicht”, 1989-1991, Wörterbücher: Ein internationales handbuch zur lexikographie, eds. Franz Josef Hausmann, Oskar Reichman, Herbert Ernst Wiegand, Ladislaw Zgusta. Berlin-New York, Walter de Gruyter, 3 vols: 649-57.

Khemakhem, M. *et al.* (2018), “Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories. Leveraging Open Data”. Tokio, Japón.

Nalessio, Giulia (2024), “Humanidades digitales y lexicografía bilingüe: recuperación y valorización del patrimonio lexicográfico español-italiano (REVALSI)”, *Lexicografía del español: internacionalización y contrastes*, eds. Alejandro Fajardo Aguirre; Dolores Torres Medina; Cristian Díaz Rodríguez. Frankfurt am Main, Peter Lang: 117-34.

San Vicente, Félix ed. (2008-2010), *Textos Fundamentales de la lexicografía italoespañola (1917-2007) / (1805-1916)*, Monza, Polimetrica.

Terras, Melissa (2010), “The Rise of Digitization: An Overview”, *Digital Libraries*, ed. Rico Rukowski. Países Bajos, Sense Publishers: 3–20.

Valero Gisbert, María J. (2016), “Fraseología y marcas diacrónicas en la lexicografía bilingüe de los diccionarios más representativos de español/italiano del s. XXI”, *El diccionario en la encrucijada: de la sintaxis y la cultura al desafío digital*, eds. Ignacio, Sariego López; Cecilio, Garriga Santander, AELEX: 395-411.

Valero Gisbert, María J. (2018), “Las marcas cronológicas en la lexicografía bilingüe italoespañola del XIX y XX”, *Quaderns d’Italia*, 23: 131-44.

Valero Gisbert, María J. (2020), “Fraseología y marcas diatópicas del español en la lexicografía bilingüe italoespañola”, *ELUA: Fraseología y variaciones (socio)lingüísticas y diatópicas*, eds. Pedro Mogorrón Huerta, Anália Cuadrado Rey, Anexo 7: 161-79.

**María Felisa Bermejo Calleja**, doctora en Filología Romanza e Linguistica Generale (Università di Perugia), fue profesora titular de Lingua, Traduzione e Linguistica Spagnola, Departamento de Lingue e Letterature straniere e culture moderne de la Università degli Studi di Torino. Entre sus áreas de investigación se pueden distinguir las siguientes: gramática de la lengua española; estudio contrastivo de la lengua española e italiana; metalexicografía bilingüe italiano-español.

**Giulia Lanteri**, tras una licenciatura en Ciencias de la Mediación Lingüística y un máster en Lenguas Extranjeras para la Comunicación Lingüística, prosigue sus estudios lingüísticos centrándose en la lexicografía española. Actualmente es estudiante de doctorado en Humanidades Digitales en las Universidades de Turín y Génova con un proyecto dedicado al estudio del lenguaje juvenil español en las redes sociales.

[giulia.lanteri@hotmail.com](mailto:giulia.lanteri@hotmail.com)

**María Joaquina Valero Gisbert** es doctora en Lingüística por la Universidad de Valencia (España). Profesora titular de Lingua, Traduzione e Linguistica Spagnola en la Universidad de Parma, Departamento de Discipline Umanistiche, Sociali e delle Imprese Culturali, unità di Lingue e Letterature Straniere. Sus áreas de investigación son la lexicografía, la fraseología hispánica, la traducción audiovisual y la audiodescripción. Dirige el máster oficial 'MTAV' en Traducción Audiovisual (<https://elly2022.postlaurea.unipr.it/course/index.php?categoryid=27>).

[mariajoaquina.valerogisbert@unipr.it](mailto:mariajoaquina.valerogisbert@unipr.it)

**Maria Carmela Zaccone** es doctora en Digital Humanities en las Universidades de Génova y Turín. Actualmente es profesora contratada de Lengua Española en la Universidad de Milán y la de Turín y post-doc en lexicografía y fraseología en la Universidad de Turín. Su investigación se sitúa en el campo de la fraseología española e italiana y la fraseodidáctica digital. Su experiencia en el campo de la didáctica le ha permitido explorar las nuevas tecnologías utilizadas en la enseñanza ELE. Ha participado en varios congresos de carácter general y específico de su campo de investigación.

[mariacarmela.zaccone@unito.it](mailto:mariacarmela.zaccone@unito.it)