Teoría y práctica de la retrodigitalización de diccionarios: el caso del *Vocabulario de las dos lenguas toscana y castellana**

GIULIA NALESSO
Università degli Studi di Padova

Resumen
Este trabajo se enmarca en *Un nuevo espacio digital para el patrimonio lexicográfico: el "Tesoro digital de la lexicografía bilingüe español-italiano"*, un proyecto de historiografía lexicográfica que tiene como objetivo el desarrollo de un protocolo para la retrodigitalización de un conjunto de diccionarios bilingües español-italiano desde las primeras publicaciones de las que tenemos constancia hasta la mitad del siglo XX, que se harán accesibles digitalmente en un tesoro lexicográfico (TELEI, Tesoro lexicográfico español-italiano). Ilustramos el marco metodológico para la retrodigitalización del corpus, esto es, los pasos necesarios para la creación de un modelo que permitirá el entrenamiento de los programas informáticos y de una ontología del conocimiento contenido en los textos y relacionado con ellos. El resultado será un tesoro en línea de acceso abierto que proporcionará una edición digital de los textos con posibilidad de búsqueda interactiva. Presentamos, además, el primer objeto de estudio, el *Vocabulario de las dos lenguas toscana y castellana* de Cristóbal de Las Casas, cuya *editio princeps* apareció en Sevilla en 1570.

Palabras clave: humanidades digitales; retrodigitalizar diccionarios; historiografía; lexicografía; TEI.

Abstract
This paper is framed in *A new digital space for the lexicographical heritage: The "Tesoro digitale della lessicografia bilingue spagnolo-italiano"*, a lexicographic historiography project aimed at developing a protocol for the retro-digitization of a corpus of Spanish-Italian bilingual dictionaries published since their origins as far as we know until the mid-20th century, which we will make digitally accessible. It focuses on illustrating the methodological framework for the retro-digitization of the corpus, that is, the necessary steps for the creation of a model that will allow the training of the software and of an ontology of the knowledge contained in the texts and related to them. The result will be an open access online lexicographic lexicon (TELEI, *Tesoro lexicográfico español-italiano*) providing a scholarly digital edition of the texts with interactive search possibilities. We present the first object of study, the *Vocabulario de las dos lenguas toscana y castellana* by Cristóbal de Las Casas, whose first edition appeared in Seville in 1570.

Keywords: digital humanities; retro-digitization of dictionaries; historiography; lexicography; TEI.

* Este trabajo se inscribe en el marco del proyecto financiado por el MUR, Ministero dell''Università e della Ricerca italiano, *A new digital space for the lexicographical heritage: The "Tesoro digitale della lessicografia bilingue spagnolo-italiano"* (TELEI), cuya investigadora principal es C. Castillo Peña de la Universidad de Padua, a la que agradecemos por su apoyo. Forman parte del grupo de investigación las universidades de (i) Bolonia – H. Lombardini (AI), N. Peñín –, (ii) Génova – A. L. de Hériz (AI), G. Esposito y M. C. Zaccone –, (iii) Padua – C. Castillo Peña (AI) y G. Nalesso –, (iv) Pisa – E. Carpi (AI), R. M. García Jiménez y E. Pérez Vázquez –, (v) Turín – F. Bermejo (AI), A. Bori y M. Valero –, (vi) Verona – M. De Beni (AI), F. Dalle Pezze, D. Hourani Martín, E. Sartor, A. La Manna y A. Alemany Martínez – (PRIN 2022 MUR 20229W73WR – CUP C53D23004010006). El trabajo de retrodigitalización empezó ya gracias al proyecto REVALSI, acrónimo italiano de *Recupero e valorizzazione del patrimonio lessicografico spagnolo-italiano*, cofinanciado por la Universidad de Padua y el Fondo Sociale Europeo REACT EU - Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 del MUR – decreto n. 1062 del 10 de agosto de 2021 (supervisora C. Castillo Peña).

Recibido el 12/03/2024 ✦ Aceptado el 23/04/2024

1. INTRODUCCIÓN

Antes de empezar, nos parece un paso obligado esclarecer la terminología que utilizaremos en relación con los procesos de digitalización a partir de los términos ingleses *digitalization* y *digitization* que han llevado a la acuñación del equivalente 'digitización', actualmente no reconocido en español, pero empleado en ciertos ámbitos —sobre todo economía, geografía y música— para describir los procesos de transformación de lo analógico (textos, imágenes, sonidos, etc.) a lo digital[1].

Según el diccionario del inglés Collins:

- *to digitize* significa "to put information into the form of a series of the numbers 0 and 1, usually so that it can be understood and used by a computer", cuyos sustantivos derivados son *digitization* y *digitalization*;
- *to digitalize* significa "to change something such as a document to a digital form (= a form that can be stored and read by computers)" y también "to start to use digital technology such as computers and the internet to do something", cuyo derivado es *digitalization*.

Parece que valerse de recursos para 'digitizar' no es suficiente para la digitalización de procesos y aún menos para la transformación digital de objetos analógicos. En efecto, la habilitación digital consiste en codificar un dato analógico con ceros y unos, de manera que sea legible, procesable, almacenable y transmisible por un sistema informático. El *Diccionario de la lengua española* (DLE) proporciona una definición similar para la voz 'digitalización', según el cual el verbo 'digitalizar' indica:

> 1. tr. Registrar datos en forma digital.
> 2. tr. Convertir o codificar en números dígitos datos o informaciones de carácter continuo, como una imagen fotográfica, un documento o un libro.

El anglicismo 'digitización' resulta superfluo debido a que ya disponemos de un término que define los métodos de conversión tecnológica. Pese a que en Nalesso (en prensa) distinguíamos entre 'digitalización' y 'digitización': "La primera corresponde, en nuestra interpretación, a la distribución de imágenes escaneadas de la fuente en papel. La segunda, al contrario, es la transformación de un texto analógico en un texto electrónico en formato *mrf*" (Bazzaco, 2020: 536), hoy tendemos a pensar que el préstamo es un neologismo innecesario por lo que hemos optado por descartar 'digitización' y, en consecuencia, también 'digitizar' y 'digitizado'.

Por otra parte, el adjetivo 'digital' aplicado al texto lexicográfico indica que se trata de una obra nacida digital, esto es, un diccionario en línea o electrónico, como por ejemplo el DiCE – *Diccionario de Colocaciones del Español* (http://www.dicesp.com/paginas), mientras que 'digitalizado' es polisémico, ya que se refiere a (i) la visualización en soporte informático de un objeto textual impreso, a saber, un diccionario analógico escaneado o fotografiado para su consulta en un ordenador, en internet, en libro electrónico, etc.; (ii) una obra impresa que, tras un proceso de digitalización, pasa a ser un objeto digital estructurado tras una codificación semántica, esto es, se pasa del papel a la creación de un documento legible por máquina (MRF).

[1] Cabe señalar que el monolingüismo de las publicaciones científicas anglófonas ha surtido un efecto en el ámbito de las humanidades digitales, ya que se halla un uso masivo del inglés como *lingua franca* (Isasi Velasco y del Rio Riande, 2022). Sin embargo, esto no debería suponer necesariamente comunicar en una *lingua unica* (Balula y Leão, 2019: 8). Hay que tener en cuenta todo esto a fin de encontrar sistemas de comunicación alternativos y establecer una terminología adecuada para impulsar la representación del multilingüismo de la comunidad científica.

DOI: http://dx.doi.org/10.13135/1594-378X/9988
Artifara, ISSN: 1594-378X

En este segundo caso, la acción de digitalizar se refiere a la aplicación de herramientas integradas que facilitan el acceso y uso compartido de los datos según los principios FAIR de encontrabilidad, accesibilidad, interoperabilidad y reutilización (Wilkinson et al., 2016). En suma, la digitalización proporciona distintos objetos que se han sometido a procesos diferentes de transformación digital.

A este propósito, según Castillo Peña (2020: 203-204) digital y digitalizado se refieren a tres tipos de productos en el campo de la historiografía lingüística y de la filología:

1. la reproducción facsimilar digital (o digitalizada) del texto o la difusión en soporte informático para la publicación de un documento concebido en principio para ser impreso que permite una interacción mínima con el lector (escaneo de textos, versión electrónica de libros, etc.);
2. la edición crítica, o académica, digital elaborada con recursos informáticos, cuyo resultado no es necesariamente un texto interactivo (del inglés *scholarly digital edition*);
3. la edición de obras antiguas realizada con la aplicación de protocolos estándar de codificación XML, que permite la marcación semántica del documento en formato MRF ofreciendo la máxima interoperabilidad al lector/usuario, que puede consultar datos y metadatos relacionados con el texto (objetos multi-textuales interactivos).

Al trabajar con obras anteriores a la eclosión de la lexicografía electrónica, cabe añadir una especificación más: 'retrodigitalización'. Este término se ha acuñado basándose en proyectos europeos de gran extensión como DARIAH y ELEXIS que, entre otras actividades ligadas a la conservación y valorización de diccionarios antiguos, tienen el objetivo de convertirlos en objetos digitales estructurados, invirtiendo en investigaciones como *Encoding Latin-Bulgarian Dictionary*, finalizada a retrodigitalizar el *Latin-Bulgarian Dictionary* (Mihail Voynov, Alexandar Milev et al., 1945). Otro proyecto dedicado a la retrodigitalización de diccionarios es GROBID-Dictionaries (Khemakhem, Foppiano y Romary, 2017). En esta misma línea se ha trabajado con el *Foclóir Gaeilge-Béarla* (Ó Dónaill, 1977) y el *English-Irish Dictionary* (Bhaldraithe, 1959) —proyecto *New English-Irish Dictionary* (https://www.teanglann.ie/ga/); el *Diccionario da Lingua Portugueza* (Morais Silva, 1789, 1813, 1823) —proyecto MORDigital; el *Orotariko Euskal Hiztegia* (Mitxelena y Sarasola, 1986) —proyecto GROBID-Dictionaries: *Experiments with the General Basque Dictionary* (OEH) PDF (https://digilex.hypotheses.org/250). En definitiva, la lexicografía que se mueve en el marco de las humanidades digitales (HD) se ocupa de retrodigitalización al convertir en un formato digital legible por ordenador una obra publicada en papel antes de que los sistemas informáticos se introdujeran en la labor de los humanistas. En virtud de las acepciones del DLE presentadas, este procedimiento coincide tanto con la conversión de la fuente primaria en un documento digital (.pdf, .jpeg, etc.) —esto es, la adquisición de imágenes MRF—, como con la sucesiva codificación semántica del diccionario, es decir la estructuración de sus contenidos en una base de datos que identifica como tales cada uno de los elementos de su micro y macroestructura. Todo ello permite la interoperabilidad entre textos distintos y la consulta de las obras como hipertexto. Como sostiene Castillo Peña (2020: 205),

> estamos de acuerdo con Rojas Castro (2017: 6) en puntualizar que lo que distingue una edición digital no consiste tanto en la prodigalidad documental que se ofrece al lector, sino en su interactividad, es decir en el hecho de que se "pueda navegar, seleccionar, filtrar o visualizar la información estructurada —las divisiones del texto, la foliación, las intervenciones editoriales, la ortografía del original, el texto modernizado, las variantes o las notas— en función de los intereses o expectativas del usuario".

Ahora bien, el proyecto *Un nuevo espacio digital para el patrimonio lexicográfico: el "Tesoro digital de la lexicografía bilingüe español-italiano"* tiene el objetivo de producir y publicar versiones digitales de alta calidad de obras lexicográficas aparecidas desde el siglo XVI hasta la mitad del XX, que permitan una búsqueda interoperable con la finalidad de preservar este patrimonio en un portal en línea de acceso abierto. Se contribuirá a una reinterpretación de estos textos con nuevos datos y paradigmas analíticos, concibiendo este material como un producto sociohistórico y, como tal, contextualizado en el espacio y el tiempo de una comunidad. Esto llevará también a conocer mejor las dos lenguas, analizar la evolución de sus palabras y profundizar en el conocimiento de sus vocabularios, también en un eje contrastivo.

La finalidad es constituir el TELEI siguiendo la tradición del *Tesoro Lexicográfico* de Gili Gaya (1957), del *Nuevo tesoro lexicográfico de la lengua española* (NTLLE) de la Real Academia Española, del *Nuevo Tesoro Lexicográfico del Español (S. XIV-1726)* de Nieto Jiménez y Alvar Ezquerra (2007), del *Tesoro Lexicográfico del Español en América* (TLEAM) dirigido por Corbella Díaz y de la *Opera del Vocabolario Italiano* (OVI) de la *Accademia de la Crusca*. En esta óptica, se producirá un diccionario de diccionarios formado por ediciones facsimilares digitales de aquellos textos que han ido recogiendo, definiendo y consolidando la lexicografía bilingüe italoespañola. De este modo, gracias a la aplicación de las HD, se logrará reunir una gran cantidad de datos lexicográficos que ninguna biblioteca física puede custodiar de forma conjunta, ofreciendo la posibilidad de consulta simultánea de distintas obras a la vez (Tramullas, 2002).

El proceso completo consiste en la conversión de los datos textuales en datos estructurados gracias a la transformación de imágenes escaneadas en texto plano, la edición del texto plano y su codificación en formato XML: esto posibilitará la preservación de esas obras mediante el planteamiento de modelos de reconocimiento automático supervisados. Sentaremos así las bases para la construcción de un tesoro lexicográfico donde la búsqueda de una entrada en la lengua A dará lugar a las equivalencias en la lengua B, proporcionando además datos como la indicación del título del diccionario, el autor, la fecha de publicación, las variantes de un lema, etc. Se propone, por tanto, un recurso innovador, ya que hasta la fecha no tenemos constancia de una herramienta análoga en la lexicografía bilingüe para ningún par de lenguas occidentales. En concreto, se creará la infraestructura necesaria para la gestión de datos digitales que agrupe herramientas integradas de modo que se faciliten el acceso, uso compartido y reutilización de tales datos según los principios FAIR mencionados arriba.

A continuación, presentaremos la metodología de investigación que se ha elegido, basada en el análisis integral de las unidades léxicas[2] que lleva a la creación de ontologías que representen adecuadamente los datos léxicos, además de hacerlos accesibles y reutilizables. Seguidamente, expondremos un esquema general del TELEI, centrándonos en la importancia de utilizar normas y formatos orientados a la interoperabilidad y accesibilidad de los datos.

2. FUNDAMENTACIÓN METODOLÓGICA

2.1 El corpus

En lo que atañe a la selección de los productos lexicográficos que servirán para el diseño del corpus de trabajo, debería incluir diccionarios, nomenclaturas y todo tipo de glosario, aunque, en una primera fase del proyecto se trabaja con un repertorio reducido constituido por un número limitado de textos, cuya función es identificar problemas filológicos y ecdóticos.

[2] Realizado después de un atento estudio historiográfico, catalográfico y lingüístico que permite la contextualización del texto y la descripción de sus características macro y microestructurales necesarias para la retrodigitalización.

DOI: http://dx.doi.org/10.13135/1594-378X/9988
Artifara, ISSN: 1594-378X

Los criterios que rigen la recopilación del corpus se basan en los siguientes parámetros:

- cronológico, el corpus contiene al menos un diccionario por cada siglo, desde el XVI hasta mediados del XX;
- tipológico, el corpus contiene al menos uno de los siguientes diccionarios (i) alfabético y no alfabético, (ii) general y especializado, (iii) monolingüe y multilingüe;
- historiográfico, el corpus, además de textos representativos e influyentes en la historia de la lexicografía, también contiene diccionarios menos conocidos[3];
- metodológico, el corpus contiene obras que plantean problemas específicos para la retrodigitalización, como sus características tipográficas, su complejidad micro y macroestructural, su historia editorial[4].

Con el objetivo de ilustrar el flujo de trabajo, hemos elegido una obra representativa, el *Vocabulario de las dos lenguas toscana y castellana* de Cristóbal de Las Casas (1570), de la cual introduciremos, a modo de ejemplo, un análisis exploratorio presentando datos útiles para el planteamiento de buenas prácticas para la retrodigitalización.

2.2 La retrodigitalización de diccionarios

Como ya se ha explicitado, nuestro objetivo primario es la retrodigitalización de obras italoespañolas publicadas desde el siglo XVI hasta la mitad del XX, siendo un proyecto que intenta promover la recuperación y la revalorización de este patrimonio lexicográfico a través de una base de datos en línea según los principios FAIR, gracias al uso de un estándar para la edición de las obras en forma digital, la *Text Encoding Initiative* (TEI). Se trata de un esquema de directrices, las TEI *Guidelines*, para la codificación y el intercambio de textos electrónicos: un estándar de marcado que permite la etiquetación semántica de documentos, esto es, establece el significado de cada uno de sus elementos y atributos[5]. La retrodigitalización de un corpus tan amplio y heterogéneo representa un desafío científico y digital que consigue solucionar dificultades metodológicas y teóricas para las cuales la comunidad científica ha desarrollado herramientas y estándares para la modelización y codificación de documentos (Costa et al., 2021). Para ello, diseñamos un flujo de trabajo que pueda ser replicado y aplicado a todo el corpus, utilizando instrumentos que permitan la extracción automática del contenido lexicográfico y de los paratextos según las fases enumeradas abajo.

[3] Asumiendo los principios metodológicos para el desarrollo de la historiografía lingüística formulados por Gómez Asencio (2007), según el cual podemos reconstruir la historia de la difusión de las lenguas si consideramos la mayor parte de los textos que lo hicieron posible, no solo los más prestigiosos o conocidos.

[4] Nos referimos en particular a: (i) tipografía, uso de diferentes tipos de caracteres, estructuras de párrafo y columnas para establecer la macroestructura del diccionario y distinguir la lengua de partida de la de llegada en la microestructura; (ii) puntuación diacrítica para distinguir distintos tipos de información lexicográfica, como equivalencias, definiciones, abreviaturas, información gramatical, ejemplos; (iii) variantes ortográficas, errores, hápax, formas no canónicas, etc. resultantes de errores de impresión o del conocimiento imperfecto de la lengua por parte del autor; (iv) variantes y cambios debido a ediciones posteriores a la primera, a veces muy numerosas o publicadas durante periodos de tiempo muy largos.

[5] La TEI fue creada con el propósito de superar la gran proliferación de esquemas de codificación incompatibles entre sí que dificultaban la investigación científica a través de distintas aplicaciones informáticas desarrolladas para los textos digitales, especialmente en humanidades. Sirve para representar (i) la estructura abstracta de diversos tipos de texto (prosa, poesía, teatro, manuscritos, fichas, material lexicográfico, etc.); (ii) las características textuales relevantes para diferentes ámbitos de investigación (filología, análisis lingüístico, análisis literario etc.); (iii) otros tipos de información (como imágenes y sonidos). Estas pautas permiten trabajar con más de 500 elementos textuales agrupados en veinte módulos que se reparten en generales, o básicos, para la marcación de cualquier documento, y especializados, para trabajar con tipologías definidas de textos, entre los que contamos con un modelo para el etiquetado de diccionarios, TEI *Dictionaries* (para más información consúltese https://tei-c.org/release/doc/tei-p5-doc/es/html/index.html) y TEI Lex-0 en su versión actualizada.

2.2.1 Flujo de trabajo

En detalle, los pasos que hemos planteado son los siguientes:

- Fase 0: Preproducción.
  - Búsqueda del material para la recopilación del corpus.
  - Análisis y creación del esquema de retrodigitalización:
    - estudio filológico e historiográfico;
    - establecimiento de los criterios de transcripción y edición, modelización, lenguaje de marcado.
- Fases 1 y 2: Precodificación.
  - Conversión de la fuente primaria en documento digitalizado (.pdf, .jpeg, .png, .tiff, etc.).
  - Transcripción de los textos digitalizados:
    - opción 1: transcripción manual de los textos digitalizados;
    - opción 2: transcripción automatizada de los textos digitalizados (OCR/HTR):
      - entrenamiento de un modelo: eliminación del ruido, segmentación de la página, resolución de los problemas, control del resultado;
      - aplicación del modelo y transformación en textos electrónicos (.txt, .doc, .pdf, etc.), esto es, en documentos MRF.
- Fases 3 y 4: Codificación y modelización.
  - Etiquetación semántica XML/TEI, conversión de datos textuales en datos estructurados:
    - identificación de elementos y atributos para la interoperabilidad y consulta entre textos como hipertexto.
  - Organización del documento en cascada (segmentación).
  - Aplicación de hojas de estilo para la visualización del texto digital.
- Fase 5: Postproducción.
  - Publicación e implementación del portal con los textos, accesibles en abierto e interoperables (FAIR).

En la fase inicial (0), que hemos llamado preproducción, se analiza el corpus, esto es, se realiza un estudio ecdótico, historiográfico y lexicográfico de las obras y, a continuación, se procede a la creación de un esquema de retrodigitalización basado en el establecimiento y la aplicación de criterios de transcripción y codificación. Seguidamente, la precodificación (fases 1 y 2) es la conversión de la fuente primaria en documento digital que facilita las operaciones de transcripción automatizada, la cual proporciona documentos MRF listos para la codificación semántica (fase 3), que se realiza a través de un análisis sintáctico para la extracción automática del contenido. Esto es posible gracias a una organización del documento en cascada (o árbol) que prevé la segmentación de:

- Diccionario.
  - Paratextos.
  - Cuerpo del diccionario.
    - Artículos.
      - ◊ Lemas y entradas.
        - Categoría de información (significado, equivalencia, información gramatical, registro u otra marcación de uso, etc.).

DOI: http://dx.doi.org/10.13135/1594-378X/9988
Artifara, ISSN: 1594-378X

Sigue la conversión de los datos textuales en datos estructurados que identifican elementos y atributos para la interoperabilidad y consulta de hipertextos. Se pasa, a continuación, a la aplicación de hojas de estilo para la visualización del texto digital (fase 4). Por último, (fase 5), la postproducción, coincide con el control de los resultados, la publicación e implementación del portal con los textos retrodigitalizados, accesibles en abierto y consultables por lema, contenido lexicográfico, contexto, etc. (por etiquetas y por relaciones léxicas), gracias al uso de datos enlazados.

2.2.2 Criterios de transcripción

Tras la fase de preproducción, el punto de partida del proceso de retrodigitalización coincide con el establecimiento de criterios de transcripción. Se trata de un paso fundamental, ya que servirá también como guía para la propuesta de futuras ediciones críticas: en estos criterios se basará el quehacer del editor que utilizará el texto transcrito y digitalizado. Para ello, es posible optar por la aplicación (i) de criterios editoriales conservadores que garanticen el respeto por el estado de la lengua del autor y/o (ii) de criterios modernizadores, esto es, intervenciones que mejoren la cohesión y lectura de la obra.
Según nuestro propósito de poner al alcance del lector moderno un texto coherente que al mismo tiempo reproduzca las características del original, se aplicarán ambas tipologías de criterio limitando las intervenciones editoriales a la unificación de cuestiones tipográficas y de ciertas irregularidades de la técnica lexicográfica para que la versión retrodigitalizada respete fielmente la fuente primaria. En el momento en que se redacta este trabajo no se han fijado de forma definitiva estos patrones, sin embargo, podemos señalar que abordarán cuestiones de:

- aparición de espacios dobles y triples;
- desajustes en el orden alfabético;
- irregularidad en el uso de las sangrías;
- irregularidad en el uso de los signos de puntuación y otros signos gráficos con valor metalexicográfico (separación de acepciones, de equivalentes, de ejemplos);
- uso de la letra cursiva y redonda;
- uso de abreviaturas;
- uso de mayúsculas y minúsculas;
- variedad en la visualización de los artículos.

2.2.3 Transcripción y codificación

La fase sucesiva consiste en la trascripción manual del texto o la aplicación de un programa de transcripción automatizada. Para la segunda opción, elegimos *Transkribus*[6], un sistema HTR, que transcribe automáticamente textos digitalizados partiendo de imágenes en diferentes formatos y extensiones (.jpeg, .png, .bitmap, .tiff, etc.) y los convierte en textos electrónicos (.txt, .doc, .pdf, etc.) para la codificación[7] con *Oxygen*[8], que edita y procesa archivos .xml. Codificar

[6] *Transkribus* (https://readcoop.eu/transkribus/) fue desarrollado por la Universidad de Innsbruck en colaboración con grupos de investigación europeos en el marco del proyecto READ (*Retrieval and Entrichment of Archival Documents*) financiado por el plan Horizon2020 en 2015.
[7] La codificación es una actividad científica, no existe un marcado neutro porque se trata de una interpretación del investigador que trabaja con el texto. Es este un proceso que puede apoyar la investigación, pero también ser un motivo de investigación por lo que es necesario un análisis del documento antes de la etiquetación para que el marcado resultante sea válido.
[8] Este programa presenta una interfaz de fácil manejo con un gran número de funciones de edición XML, permite corroborar la corrección formal de los marcados y, además, ofrece la posibilidad de producir una serie de documentos, preestablecidos y configurables, en diferentes formatos para su publicación (https://www.oxygenxml.com).

un texto significa traducirlo para que la máquina pueda leerlo cumpliendo los siguientes pasos: análisis preliminar del texto; transformación en versión digital; establecimiento de un lenguaje de marcado y de un esquema de codificación; conversión del texto MRF y etiquetación; visualización en línea; control de los resultados y publicación.

Cabe señalar que los resultados de la transcripción automatizada dependen de la calidad del documento-fuente. A este respecto, se deben adquirir imágenes de alta resolución para procesarlas con el mínimo margen de error; en particular este impacto afecta aún más al resultado a la hora de extraer datos granulares (Khemakhem et al., 2019; Bazzaco, 2020; Nalesso, en prensa), que para TELEI serían los niveles menores de segmentación del diccionario. De hecho, el cuerpo de un documento .xml está estructurado de forma segmentada y se compone de elementos organizados necesariamente en una estructura de árbol con un elemento raíz. Para que este proceso tenga éxito, es preciso adquirir correctamente el texto-imagen conforme a las siguientes etapas:

- adquisición de la imagen de la fuente primaria (analógica);
- eliminación del ruido, aquellos rasgos del texto digitalizado que pueden causar errores de reconocimiento, por ejemplo, sería deseable utilizar imágenes en colores y limpiarlas eliminando eventuales manchas y defectos;
- segmentación de la página que permite el correcto orden de lectura automática (*Layout Analysis*);
- entrenamiento de un modelo *ad hoc* que sepa reconocer un tipo de letra específico, una determinada disposición del texto, etc. a través de la transcripción manual de una pequeña porción de este texto, tras fijar los criterios de transcripción (*Ground Truth*);
- aplicación de medidas de control de calidad, es decir averiguación de los resultados y del margen de error mediante un valor porcentual definido CER (*Character Error Rate*), que prueba la distancia entre texto reconocido y texto original[9].

Se obtendrá así un texto exportable en el formato requerido (.txt, .doc, .pdf, etc.) para su codificación. A estas alturas, será posible etiquetar los contenidos que forman la macro y microestructura del diccionario en elementos y atributos. De ahí, tendremos una metodología replicable sistemáticamente para las obras del corpus[10] en el portal donde será posible buscar datos y metadatos. La reestructuración de los documentos fuente en documentos codificados se desarrolla a través de la aplicación de estándares que representan la estructura jerárquica, que según las directrices TEI está formada por los siguientes niveles:

- <entry>, entrada estructurada;
- <entryFree>, entrada no estructurada, que contiene una entrada no necesariamente conforme a las restricciones impuestas por el elemento <entry>;
- <hom>, información sobre homógrafos;
- <sense>, información sobre el significado del lema;
- <dictScrap>, parte de la entrada del diccionario en la que otros elementos de nivel sintagmático del diccionario se combinan de forma libre[11].

[9] Asimismo, conviene tener en cuenta y solucionar los posibles problemas derivados de la transcripción automática: por ejemplo, uso alternado de dos lenguas, abreviaturas no codificadas, variantes, imprecisiones técnicas, signos de difícil interpretación, organización y color de la página, deformaciones del documento.

[10] Según TEI, "[…] la estructura de las entradas de los diccionarios varía mucho entre obras distintas, la forma más sencilla para que un esquema de codificación se adapte a toda la gama de estructuras que se encuentran en la actualidad es permitir que cualquier elemento aparezca en cualquier lugar de una entrada." (La traducción es nuestra, https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html).

[11] Puede contener cualquier elemento del diccionario en cualquier combinación para dar información de nivel inferior, que no forma por sí misma una unidad estructural identificable.

DOI: http://dx.doi.org/10.13135/1594-378X/9988
Artifara, ISSN: 1594-378X

Estos pueden contener uno o varios elementos, de nivel superior, con datos relacionados con la información sobre la forma del lema (ortografía, pronunciación, separación silábica, etc.), información gramatical (parte de la oración, subcategorización gramatical, etc.), definiciones o traducciones a otra lengua, etimología, ejemplos, información de uso, referencias cruzadas a otras entradas, notas (metadatos del documento digital y de su fuente analógica, por ejemplo, referencias a personas, instituciones, lugares, fechas, acontecimientos, etc.), entradas/palabras (a menudo de forma reducida) relacionadas con otros lemas.

Los constituyentes de nivel superior se codifican a través de otros elementos y atributos inferiores, que pueden incluir:

- <form>, información sobre la forma oral y escrita del lema;
- <gramGrp>, información morfosintáctica;
- <def>, definición;
- <cit>, cita a una referencia bibliográfica;
- <usg>, información de uso;
- <xr>, reenvío hacia otros puntos de este u otro texto;
- <etym>, información etimológica;
- <re>, entrada que se incluye en una entrada mayor para un elemento léxico relativo al lema, como un sintagma compuesto o una forma derivada;
- <note>, nota o aclaración.

El resultado será un tesoro lexicográfico histórico que trasciende el concepto tradicional de diccionario, ajustándose a la e-lexicografía y siguiendo las labores de los mencionados NTLLE, TLEAM y OVI. Los textos serán accesibles en abierto y se podrán buscar a través de una interfaz interactiva.

3. ESTUDIO DE CASO: LA RETRODIGITALIZACIÓN DEL *VOCABULARIO DE LAS DOS LENGUAS TOSCANA Y CASTELLANA*

3.1 Advertencia

En este apartado proporcionamos un ejemplo práctico del desarrollo de las primeras fases del flujo de trabajo, que hasta la fecha hemos aplicado a la edición príncipe del *Vocabulario de las dos lenguas toscana y castellana* de Cristóbal de Las Casas (1570), dando cuenta del estado de la cuestión actual. Nos centraremos en una parte de la preproducción, que coincide con el estudio historiográfico del texto para reflexionar sobre ciertas cuestiones que constituyen desafíos para la creación del esquema de retrodigitalización. Dejamos de lado las fases 1 y 2 de precodificación, detalladas en Nalesso (en prensa), para pasar a explicar brevemente las demás etapas.

3.2 Apuntes historiográficos

Podemos caracterizar la obra en virtud de su tipología, destinatario y función perseguida según el patrón de planificación lexicográfica TDF (Matus, 2007 *apud* Chávez Fajardo, 2022: 28). Se trata de un diccionario bilingüe y semasiológico, que cataloga el léxico por orden alfabético en ambas secciones (una en la dirección ITA>ESP y la otra ESP>ITA), que presenta un lemario formado por una entrada en la lengua de partida con su traducción a la lengua de llegada.

La portada[12] introduce los siguientes datos: título, autor, destinatario, fecha y lugar de publicación. Además, declara la presencia de una sección dedicada al aprendizaje de las reglas de pronunciación de los dos idiomas. El metalenguaje es el español escrito en redonda. Por otra parte, las muestras en lengua italiana están en cursiva: "Sacanse las diciones, que en latin se escriven con *l* liquida, y en Toscano se mudo la *l* en *i* como *Fiore*, en latin *Flos*, *Chiaue* […]" (Las Casas, 1570: 10r) o "La *g* con la *l* sucediendo les *i* suena entre nosotros dos *ll* como *Foglia*, *Boglire*, *Moglie*, *Boglio* […]" (Las Casas, 1570: 11r). Aparece en cursiva también el latín para explicar ciertos rasgos de las dos lenguas objeto. La estructura de los artículos resulta simple y sistemática, ya que, además de las equivalencias y esporádicos sinónimos o datos adicionales (véase §3.3), no aparece ninguna información sobre el lema (forma, sintaxis, etimología, uso) y tampoco hallamos un sistema de ejemplificación basado en un canon explícito, pues no hay citas de autoridades. Parece, en suma, un inventario normativo redactado con pretensiones didácticas entre los siglos XVI-XVII para aprender y usar correctamente los idiomas toscano y castellano como lenguas extranjeras.

En lo que a datos generales se refiere, resumimos la información catalográfica con la siguiente ficha[13].

Título	Vocabulario de las dos lenguas toscana y castellana
Tipología	Diccionario bilingüe general
Secciones de la obra	Paratextos iniciales Reglas de pronunciación Vocabulario ITA>ESP ESP>ITA Paratextos finales
Autor	Cristóbal de Las Casas
Portada	VOCABULARIO / DE LAS DOS LENGVAS TOSCA / NA Y CASTELLANA DE CHRIS- / TOVAL DE LAS CASAS. / EN QVE SE CONTIENE LA DECLARA- / *cion de Toscano en Castellano, y de Castellano / en Toscano. En dos partes* / CON VNA INTROVCION PARA LEER, / *y pronunciar bien entrambas lenguas*. / DIRIGIDO AL ILLVSTRISSIMO / *señor don Antonio de Guzman, Marques de* / *Ayamonte, señor delas villas de* / *Lepe y la Redondela*. / [adorno] / Con Priuilegio de Castilla y de Aragon. / *Vende se en casa de Francisco Aguilar mercader de libros*. / EN SEVILLA. / 1570 [*sic*.]
Edición	Primera
Año edición	1570

[12] Manejamos la obra digitalizada de la *Biblioteca Digital Hispánica* (http://bdh-rd.bne.es/viewer.vm?id=0000127110&page=1) del ejemplar U/6948 (Colección Usoz) localizado en la Sede de Recoletos de la Biblioteca Nacional de España, compuesta por 248 hojas.

[13] Nos basamos en el modelo de catalogación aplicado al portal LITIAS. *La lingua italiana in territori ispanofoni, da lingua della cultura e della traduzione a lingua dell'educazione* (http://litias.cliro.unibo.it/wp/), un proyecto de investigación PRIN (Proyectos de Investigación de Interés Nacional, protocolo 2017J7H322) financiado por el MUR italiano. Cada ficha presenta dos versiones: una breve con información catalográfica básica (título; género textual; autor y eventuales traductores, redactores y colaboradores; fecha de la edición o reimpresión; copyright; ISBN; editor; tipógrafo; lugar de edición; número de volúmenes; colección; número de páginas) y otra extensa en la que se añaden a los datos típicamente bibliográficos de la primera una serie de informaciones imprescindibles para los estudios historiográficos (índice; apéndices; características del léxico; notas tanto sobre la edición o reimpresión del texto catalogado como sobre sus otras ediciones o reimpresiones; bibliografía crítica; notas del traductor u otras notas importantes; localización).

DOI: http://dx.doi.org/10.13135/1594-378X/9988
Artifara, ISSN: 1594-378X

Reimpresiones y otras ediciones	12
Año reimpresiones y otras ediciones	En Venecia: 1576, 1582, 1587, 1591, 1597, 1600, 1604, 1608, 1613, 1618, 1622 En Sevilla: 1583
Derecho de autor	Firmado el 19 de agosto de 1569 por Antonio de Erasso (por mando del rey)
Editor/Impresor	Alonso Escriuano
Lugar de edición	Sevilla
Volúmenes	1
Número de páginas	248
Índice[14]	- Portada [1r] - Aprobaciones [1v] - [*Imprimatur*] El rey [2r-3r] - [Dedicatoria] Al Illvstrissimo señor don Antonio de Guzman [3v-5v] - [Composiciones poéticas] [6r-9r] - - [Carmen de Juan de Mallara] [6r-6v] - - [Epigrama de Francisco López] [6v] - - [Carta de Fernando de Herrera] [7r-7v] - - [Carta de Pedro Laínez] [8r-8v] - - [Soneto de Juan de Vadillo] [9r] - [Página en blanco] [9v] - [Cuerpo de la obra] - - [Reglas de pronunciación] Introducion para leer, y pronunciar bien las lenguas Toscana y Castellana [10r-12v] - - [Vocabulario] [13r-247r] - - [Portada] Primera parte del vocabulario de la lengva Toscana y Castellana [13r-153v] - - - [Vocabulario italiano-español] [AB.]-ZV. [13r-153v] - - [Portada] Segvnda parte del vocabulario de las dos lengvas Castellana y Toscana [154r-247r] - - - [Soneto de Juan de Mallara] [154r] - - - [Vocabulario español-italiano] AB.-ZV. [155r-247r] - Erratas [247v] - Registro [248r] - [Página en blanco] [248v]
Apéndices	Fe de erratas Colofón

Tabla 1. Ficha catalográfica del Vocabulario de las dos lenguas toscana y castellana *(1570).*

Se trata de información de interés, ya que permite cumplir uno de los primeros pasos de la preproducción para la subsiguiente organización estructurada de las obras del corpus. En efecto, el producto final de un proceso de catalogación sistematizada llevará a la materializa-

[14] Este índice razonado, recopilado según las pautas de Lombardini y San Vicente (2015), sirve para organizar el texto en función de su estructura jerárquica. Los corchetes marcan epígrafes que han sido añadidos para presentar todos los contenidos de la obra.

ción del portal para el acceso a las fichas de los textos y la consulta de su versión retrodigitalizada.

Como hemos visto en el apartado previo, la fase preliminar de la investigación consiste en el estudio filológico e historiográfico de los materiales, que es necesario para el establecimiento de los criterios de transcripción y edición, las pautas de modelización y el lenguaje de marcado. A continuación, se pasa a analizar la estructura del diccionario, ya que conocer la segmentación de la página y los elementos de la entrada permiten un correcto orden de lectura para la transcripción y etiquetación a fin de obtener la versión retrodigitalizada.

3.2.1 Ediciones y reimpresiones

Además de la *editio princeps* aparecida en Sevilla en 1570, y otra publicada en esta misma ciudad trece años después, según Gallina (1959) y Acero (1991) disponemos de once ediciones venecianas publicadas por diferentes editores (1576, 1582, 1587, 1591, 1597, 1600, 1604, 1608, 1613, 1618, 1622). Las búsquedas que realizamos en distintos repositorios en línea (Edit 16, BVFE, REBIUN, *WorldCat*) permiten confirmar la existencia de los trece ejemplares. Por su parte, Lope Blanch (1990) da cuenta de tres publicaciones sevillanas (1570, 1579, 1583) y doce venecianas (1576, 1582, 1587, 1591, 1594, 1597, 1600, 1604, 1608, 1613, 1618, 1622), entre las cuales no hemos podido encontrar las de 1579 y 1594. Niederehe (1994, 1999) añade otros testimonios que no se han localizado: 1576 (Sevilla) y 1576, 1582, 1587, 1588, 1591, 1594, 1597, 1600, 1604, 1613, 1622 (Venecia). San Vicente (2010) señala también un ejemplar veneciano, que no hallamos en bibliotecas físicas o virtuales, fechado en 1662[15].

La siguiente tabla reúne los ejemplares obtenidos de nuestras búsquedas.

Editor	Edición/ reimpresión	Fecha	Lugar
Alonso Escriuano	1ª edición	1570	Sevilla
Damian Zenaro	2ª edición	1576	Venecia
Damian Zenaro	3ª edición	1582	Venecia
Andrea Pescioni	Reimpresión[16]	1583	Sevilla
Damian Zenaro	Reimpresión[17]	1587	Venecia
Damian Zenaro	Reimpresión	1591	Venecia
Damian Zenaro	Reimpresión	1597	Venecia
Oliuier Alberti	Reimpresión	1600	Venecia
Guerra Fratelli	Reimpresión	1604	Venecia
Matthio Valentino	Reimpresión	1608	Venecia
Marc'Antonio Zaltieri	Reimpresión	1613	Venecia
Giovanni Antonio Giulani	Reimpresión	1618	Venecia
Pedro Miloco	Reimpresión	1622	Venecia

*Tabla 2. Lista de ediciones y reimpresiones del V*ocabulario de las dos lenguas toscana y castellana

A partir de esta pluralidad de ediciones y reimpresiones, tras el proceso de retrodigitalización, será posible crear una versión navegable de la obra basada en la príncipe que permita

[15] Él mismo tiene constancia bibliográfica de algunos, pero no los sitúa: 1576 (Sevilla, eds. Aguilar y Escriuano); 1588 (Venecia, ed. Zenaro); 1594 (Venecia, ed. Bertano); 1662 (Venecia, ed. Miloco).
[16] Se trata de una reimpresión de la príncipe.
[17] Se trata de una reimpresión de la edición veneciana de 1582.

DOI: http://dx.doi.org/10.13135/1594-378X/9988
Artifara, ISSN: 1594-378X

consultar los testimonios y dar cuenta de la variación micro y macroestructural del *Vocabulario*, gracias a las posibilidades de interrogación de los datos. En concreto, se documentarán los cambios (i) macroestructurales (lemario, supresión o adición de entradas y equivalencias, alteración del orden alfabético, organización del texto), (ii) microestructurales (tipo de información metalingüística y rasgos tipográficos), (iii) lingüísticos (estado de la lengua, rasgos ortográficos).

3.3 Macro y microestructura

Si excluimos los paratextos (4,44%), el cuerpo textual cubre el 95,56%. La introducción (1,21%) presenta un apartado dedicado a reglas de lectura y pronunciación de ambas lenguas, con claras finalidades didácticas, "Introducion para leer, y pronunciar bien las lenguas Toscana y Castellana" y "Advertencia en la pronunciacion Castellana". El repertorio lexicográfico (94,35%) se distribuye en una primera parte en la que aparecen los lemas italianos con su traducción al español (56,85%) y una segunda que contiene los artículos españoles con sus equivalentes italianos (37,50%), ordenados alfabéticamente y colocados en doble columna por folio:

> el alfabeto italiano es el siguiente: A, B, C, D, E, F, G, H, I, L, M, N, O, P, Q, R, S, T, V, Z. Por ende, está compuesto por 22 letras, a pesar de que en el lemario sean 20, ya que no se distingue la U mayúscula de la V y no se contemplan las letras K y X. A este propósito, "La x no se vea en esta lengua, aunque algunos la ponen en principio de nombres propios, como *Xerse*" (Las Casas 1570: 10v). Por su parte, el alfabeto español es el mismo que en italiano: A, B, C, D, E, F, G, H, I (se incluye aquí la Y, con las siguientes subdivisiones: YA, YE, YO, YV y también la J en las reglas de pronunciación que preceden el vocabulario), L (aparece también el dígrafo LL), M, N (de entre los artículos no hay una sección dedicada a la Ñ, pero en la introducción se describe como letra independiente para explicar el fonema italiano /gn/), O, P, Q, R, S, T, V (se distingue entre vocal y consonante), X, Z. (Nalesso, en prensa)

Cada letra está subdividida en secciones internas, por ejemplo, en la sección primera la A se reparte en AB-AC-AD-AE-AF-AG-AH-AI-AL-AM-AN-AP-AQ-AR-AS-AT-AV-AZ y la B en BA-BE-BI-BL-BO-BR-BV; en la segunda AB-AC-AD-AE-AF-AG-AH-AI-AL-AM-AN-AO-AP-AQ-AR-AS-AT-AV-AX-AZ y BA-BE-BI-BL-BO-BR-BV.

Si cotejamos la primera edición (sevillana) con las ediciones de 1576 y 1582 (venecianas)[18] observamos que en estas dos últimas se añaden paratextos metalingüísticos en italiano que preceden al repertorio lexicográfico después de las reglas de pronunciación, a diferencia de la príncipe en la que la lengua vehicular es solo el español: "Osservationi, ouero introduttioni della Lingua Castigliana" (5 páginas) y "Della ortografia et mvtamento di lettere della Lingua Castigliana" (19 páginas). Asimismo, desaparecen las aprobaciones, el *imprimatur* y la fe de erratas. A la dedicatoria del autor se añade una nueva, escrita por el editor Damiano Zenaro, la misma en ambas ediciones venecianas, fechada en 1576. De ahí que las nuevas ediciones presenten la siguiente subdivisión: en la segunda, los paratextos iniciales (portada, dedicatorias, composiciones poéticas) y finales (registro y marca tipográfica) cubren el 4,66%, los tratados el 6,36% y el repertorio lexicográfico el 88,98% (59,76% en la dirección ITA>ESP, pp. 1-267[19], y 40,24% en la parte ESP>ITA, pp. 269-437), por lo que el cuerpo textual se extiende

[18] Trabajamos con ediciones digitalizadas: la de 1576 de la *Biblioteca Digital Hispánica*, localizada en la Sede de Recoletos de la Biblioteca Nacional de España (signatura U/8300 – Colección Usoz), y la de 1582 de la Biblioteca Nazionale Centrale de Roma, localizada con signatura 6. 30.F.30.

[19] Se omiten los números de página de 241 a 256.

hasta el 95,34%. La configuración de la tercera es similar, ya que los elementos paratextuales llegan al 4,26%, los didácticos al 6,38% y el lemario al 89,36% (señalamos la misma repartición de las dos secciones), pues el cuerpo alcanza el 95,74%. Además, hemos cotejado también el texto veneciano de 1591[20], ya que por lo anunciado en la portada hay algunos cambios que en realidad no hemos podido detectar: parece una copia de la edición de 1582, que a su vez repite el lemario de la de 1576 (el cual presentaba ya más entradas con respecto a la de 1570). Todo esto pese a que la portada de 1582 anuncie "Et accresciuto da Camillo Camilli[21] di molti vocaboli, che non erano nella prima impressione" y la de 1591 "Et accresciuto di nuouo da Camillo Camilli di molti vocaboli, che non erano nell'altre impressioni"[22].

Un análisis macroestructural más detallado aporta otros datos cuantitativos sobre el lemario. Ante todo, la caja del texto de la príncipe, en sendas direcciones, llega a un máximo de 35 entradas por columna, por lo que contamos con aproximadamente 29.696 lemas repartidos en 17.920 ITA>ESP y 11.776 ESP>ITA. A la hora de cotejar las cuatro ediciones destacamos, a partir de la segunda, un incremento del 7,49% en el total de entradas (6,45% en ITA>ESP y 9,07% en ESP>ITA), puesto que la caja del testo permite 38 entradas por columna, lo que lleva a un total de 31.929 (19.076 ITA>ESP y 12.844 ESP>ITA). Confirmamos el desequilibrio de las dos secciones, ya señalado por Gallina (1959).

Para terminar, como ya se ha dicho, la entrada, tanto en la primera parte como en la segunda, se estructura tipográficamente igual en todas las ediciones que hemos comparado: los lemas italianos aparecen en cursiva y los españoles en redonda, se separan por puntos o por comas en el caso de los sinónimos. En lo que al tipo de información metalingüística se refiere, aparecen de vez en cuando datos sobre la categoría gramatical de la entrada o informaciones diafásicas y/o diatópicas. La entrada suele estar compuesta por un solo lema con uno o más equivalentes, sin otro tipo de información lingüística, aunque, ocasionalmente, se proporcionan más detalles como:

(1) *Ah.* Ay quexandose.
(Las Casas, 1570: 17v)

(2) *E.* Y conjuncion.
(Las Casas, 1570: 57v)

(3) *Babbo.* Boz de niño, q llama a su padre.
(Las Casas, 1570: 27r)

(4) *La.* La, articulo femenino.
(Las Casas, 1570: 84v)

(5) *Indegnità.* Indignidad.
Indegnitate. Lo mesmo.
(Las Casas, 1570: 78v)

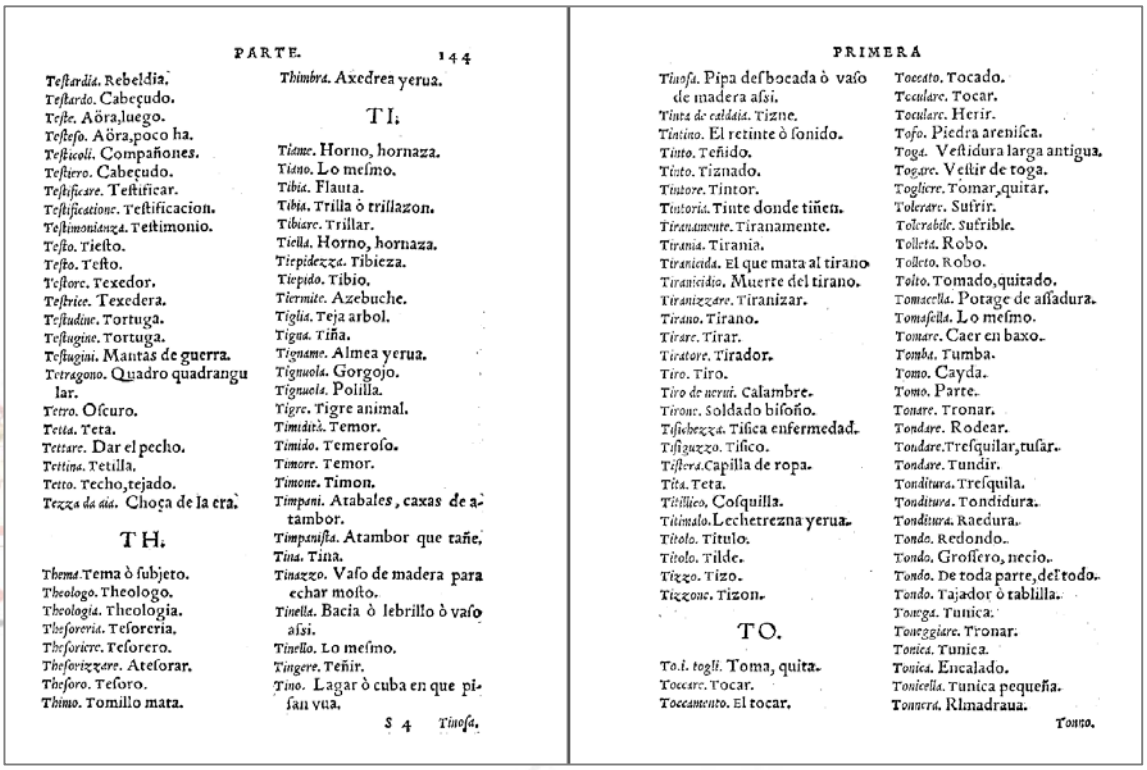
La imagen que se ofrece a continuación ilustra las características formales que presenta la obra, relacionadas con la maquetación del texto, la jerarquía interna y la división de las secciones.

[20] Manejamos el ejemplar digitalizado de la Bayerische Staatsbibliothek de Múnich, localizado con signatura L.lat.f. 309.

[21] Según EDIT 16, fue un intelectual toscano, poeta y traductor de varias obras del latín y del español al italiano. Nació posiblemente en Siena y murió en Ragusa en 1615, donde trabajó como profesor.

[22] No es objetivo de este trabajo el análisis pormenorizado del número de lemas que contiene la obra, que será necesario a la hora de crear una edición crítica del *Vocabulario*, o bien de vaciar las voces para el tesoro digital. Nuestro propósito ahora es el de dar cuenta de las líneas generales de la obra para fijar el esquema de segmentación y codificación.

DOI: http://dx.doi.org/10.13135/1594-378X/9988
Artifara, ISSN: 1594-378X



PARTE. 144

Teſtardia. Rebeldia.
Teſtardo. Cabeçudo.
Teſte. Aöra,luego.
Teſteſo. Aöra,poco ha.
Teſticoli. Compañones.
Teſtiero. Cabeçudo.
Teſtificare. Teſtificar.
Teſtificatione. Teſtificacion.
Teſtimonianza. Teſtimonio.
Teſto. Teſto.
Teſto. Teſto.
Teſtore. Texedor.
Teſtrice. Texedera.
Teſtudine. Tortuga.
Teſtugine. Tortuga.
Teſtugini. Mantas de guerra.
Tetragono. Quadro quadrangular.
Tetro. Oſcuro.
Tetta. Teta.
Tettare. Dar el pecho.
Tettina. Tetilla.
Tetto. Techo,tejado.
Tezza da aia. Choça de la era.

TH.

Thema Tema ò ſubjeto.
Theologo. Theologo.
Theologia. Theologia.
Theſoreria. Teſoreria.
Theſoriere. Teſorero.
Theſorizzare. Ateſorar.
Theſoro. Teſoro.
Thimo. Tomillo mata.
Thimbra. Axedrea yerua.

TI.

Tiame. Horno, hornaza.
Tiano. Lo meſmo.
Tibia. Flauta.
Tibia. Trilla ò trillazon.
Tibiare. Trillar.
Tiella. Horno, hornaza.
Tiepidezza. Tibieza.
Tiepido. Tibio.
Tiermite. Azebuche.
Tiglia. Teja arbol.
Tigna. Tiña.
Tigname. Almea yerua.
Tignuola. Gorgojo.
Tignuola. Polilla.
Tigre. Tigre animal.
Timidità. Temor.
Timido. Temeroſo.
Timore. Temor.
Timone. Timon.
Timpani. Atabales, caxas de atambor.
Timpaniſta. Atambor que tañe.
Tina. Tina.
Tinazzo. Vaſo de madera para echar moſto.
Tinella. Bacia ò lebrillo ò vaſo aſsi.
Tinello. Lo meſmo.
Tingere. Teñir.
Tino. Lagar ò cuba en que piſan yua.

S 4 *Tinoſa.*

PRIMERA

Tinoſa. Pipa deſboçada ò vaſo de madera aſsi.
Tinta de caldaia. Tizne.
Tintino. El retinte ò ſonido.
Tinto. Teñido.
Tinto. Tiznado.
Tintore. Tintor.
Tintoria. Tinte donde tiñen.
Tiranamente. Tiranamente.
Tirania. Tirania.
Tiranicida. El que mata al tirano
Tiranicidio. Muerte del tirano.
Tiranizzare. Tiranizar.
Tirano. Tirano.
Tirare. Tirar.
Tiratore. Tirador.
Tiro. Tiro.
Tiro de nerui. Calambre.
Tirone. Soldado biſoño.
Tiſichezza. Tiſica enfermedad.
Tiſiguzzo. Tiſico.
Tiſſera. Capilla de ropa.
Tita. Teta.
Titillico. Coſquilla.
Titimalo. Lechetrezna yerua.
Titolo. Titulo.
Titolo. Tilde.
Tizzo. Tizo.
Tizzone. Tizon.

TO.

To.i. togli. Toma, quita.
Toccare. Tocar.
Toccamento. El tocar.
Toccato. Tocado.
Toculare. Tocar.
Toculare. Herir.
Tofo. Piedra arenisca.
Toga. Veſtidura larga antigua.
Togare. Veſtir de toga.
Toglere. Tomar,quitar.
Tolerare. Sufrir.
Tolerabile. Sufrible.
Tolleta. Robo.
Tolleto. Robo.
Tolto. Tomado,quitado.
Tomacella. Potage de aſſadura.
Tomaſella. Lo meſmo.
Tomare. Caer en baxo.
Tomba. Tumba.
Tomo. Cayda.
Tomo. Parte.
Tonare. Tronar.
Tondare. Rodear.
Tondare. Treſquilar,tuſar.
Tondare. Tundir.
Tonditura. Treſquila.
Tonditura. Tondidura.
Tonditura. Raedura.
Tondo. Redondo.
Tondo. Groſſero, necio.
Tondo. De toda parte,del todo.
Tondo. Tajador ò tablilla.
Tonega. Tunica.
Toneggiare. Tronar.
Tonica. Tunica.
Tonica. Encalado.
Tonicella. Tunica pequeña.
Tonnera. Rlmadraua.

Tonto.

Imagen 1. *Vocabulario* (1570: 144r-144v), ejemplo de jerarquía interna.

3.4 Cuestiones abiertas

El estado actual del proyecto tiene como cuestión pendiente el desarrollo completo de la fase 3. Sin embargo, sabemos que, según las directrices TEI, para la codificación del texto es necesario establecer un modelo de segmentación y etiquetación semántica de sus componentes, con lo cual podemos afirmar que al trabajar con una obra sistematizada como el *Vocabulario* hay que plantear una estructura jerárquica simple formada por los siguientes elementos:

- <entry> o <entryFree>, que definen la entrada;
- <form> y <orth>, que establecen su forma escrita;
- <sense> y <def >, que aportan información sobre su significado, en este caso la traducción al español o al italiano.

A modo de ejemplo, una entrada de la sección ITA>ESP que presenta el lema italiano con dos equivalencias españolas tendrá la estructura (Las Casas, 1570: 144r):

```
<entry>
        <form>
                <orth>Teste</orth>
        </form>
        <sense n="1">
                <def>Aöra</def>
        </sense>
        <sense n="2">
                <def>luego</def>
        </sense>
</entry>
```

La fase 4, a saber, la aplicación de hojas de estilo para la visualización del texto digital, y la fase 5, la postproducción, quedan como asunto abierto desde el momento en que de entre las muchas posibilidades de TEI para la publicación de los diccionarios retrodigitalizados tenemos todavía que establecer el modelo más adecuado a los objetivos del proyecto, debido a la complejidad de la tipografía y de la estructura de las demás fuentes primarias:

- visualización tipográfica, que aporta información sobre los saltos de línea o página y otras características de la maquetación;
- visualización editorial, que permite visualizar la redacción y la puntuación del texto, la secuencia de los elementos, pero no ofrece detalles tipográficos;
- visualización léxica, que incluye la información subyacente del diccionario, sin datos sobre su forma.

4. CONCLUSIONES

Ante todo, una premisa terminológica ha aclarado el significado que para nosotros tienen los términos 'digitalizar' y 'retrodigitalización', ya que los usos de ciertos vocablos pertenecientes al ámbito de las HD procedentes del inglés nos parecen neologismos forzados en español, por lo que hemos descartado su aplicación. A continuación, se han descrito las posibilidades que los sistemas de transcripción automática aportan al perfeccionamiento de las técnicas de digitalización de textos y a su codificación gracias a un riguroso análisis que permite organizar y estructurar los componentes del diccionario, así como establecer relaciones léxicas entre elementos y atributos para la creación de un tesoro lexicográfico. La consulta de esta herramienta se realizará a través de un portal de acceso abierto, dotado de un sistema de búsqueda que no solo garantizará la información desde la perspectiva lexicográfica más tradicional, sino que ofrecerá la posibilidad de trabajar con diversos tipos de datos estructurados. Precisamente por esta razón nos referimos a nuestro tesoro como a una biblioteca digital, ya que publicaremos una colección de obras digitales en un sistema de información interactivo que permite organizar, actualizar y consultar datos y metadatos vinculados a distintos documentos según los principios FAIR. Así, confiamos en que el TELEI favorezca el planteamiento de un concepto innovador de diccionario que pueda utilizarse como nuevo recurso léxico-lingüístico gracias a las HD.

Para terminar, a partir del análisis de los artículos lexicográficos y componentes de la *editio princeps* del *Vocabulario de las dos lenguas toscana y castellana* de Las Casas (1570), hemos reseñado un proceso concreto de retrodigitalización, aplicable a un corpus de obras lexicográficas español-italiano (diccionarios, nomenclaturas y glosarios que aparecen en obras gramaticales u otros textos para el aprendizaje de una lengua extranjera). Sin embargo, como se ha demostrado tenemos aún que establecer ciertos criterios para poder desarrollar todas las fases del flujo de trabajo. Igualmente, dado que nos basamos en las directrices TEI, habrá que resolver otra cuestión abierta que atañe a la anotación de las variantes y la creación de una edición crítica de la obra, ya que, en el momento de la redacción de este trabajo, estas pautas proporcionan dos módulos separados, uno para la codificación de diccionarios y uno para la edición crítica de textos antiguos.

Bibliografía

ACERO, Isabel (1991) "Incorporaciones léxicas en el *Vocabulario de las dos lenguas toscana y castellana* de Cristóbal de las Casas," *Anuario de Estudios Filológicos* 14, págs. 7-14, en línea:

DOI: http://dx.doi.org/10.13135/1594-378X/9988
Artifara, ISSN: 1594-378X

https://dehesa.unex.es/flexpaper/template.html?path=https://dehesa.unex.es/bitstream/10662/2646/1/0210-8178_14_7.pdf#page=1 (29/01/2024).

ALONSO RAMOS, Margarita (s. *f.) Diccionario de Colocaciones del Español* (DiCE), en línea: http://www.dicesp.com/paginas (15/02/2024).

BALULA, Ana y Delfim LEÃO (2019) "Is Multilingualism Seen as Added Value in Bibliodiversity?: A Literature Review Focussed on Business and Research Contexts", *ELPUB 2019 23rd edition of the International Conference on Electronic Publishing,* Marseille, en línea https://hal.science/hal-02143195/document (10/03/2024).

BAZZACO, Stefano (2020) "Siglo de Oro: creación de un modelo HTR basado en libros de caballerías del siglo XVI en la plataforma Transkribus", *JANUS. Estudios sobre el Siglo de Oro* 9, págs. 534-561, en línea: https://www.janusdigital.es/descargar.htm?id=160 (03/03/2024).

LAS CASAS, Cristóbal de (1570) *Vocabulario de las dos lenguas toscana y castellana*, Sevilla, Alonso Escriuano.

—— (1576) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Damian Zenaro.

—— (1582) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Damian Zenaro.

—— (1583) *Vocabulario de las dos lenguas toscana y castellana*, Sevilla, Andrea Pescioni.

—— (1587) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Damian Zenaro.

—— (1591) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Damian Zenaro.

—— (1597) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Damian Zenaro.

—— (1600) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Oliuier Alberti.

—— (1604) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Guerra Fratelli.

—— (1608) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Matthio Valentino.

—— (1613) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Marc'Antonio Zaltieri.

—— (1618) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Giovanni Antonio Giulani.

—— (1622) *Vocabulario de las dos lenguas toscana y castellana*, Venecia, Pedro Miloco.

CASTILLO PEÑA, Carmen (2020) "Epigrama: Un portal para la edición digital de textos gramaticales". *Anales De Lingüística* 4, págs. 201–217, en línea: https://revistas.uncu.edu.ar/ojs3/index.php/analeslinguistica/article/view/4395 (29/01/2024).

CHÁVEZ FAJARDO, Soledad (2022) *Elementos de lexicografía hispanoamericana fundacional : acerca del Diccionario de chilenismos y de otras voces y locuciones viciosas de Manuel Antonio Román (1901-1918)*, Jaén, UJA.

COLLINS = Collins Online English Dictionary, en línea: https://www.collinsdictionary.com (15/02/2024).

COSTA, Rute 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, en línea: https://hal.inria.fr/hal-03195362v2 (15/02/2024).

DARIAH *WORKING GROUP LEXICAL RESOURCES* = *Digital Research Infrastructure for the Arts and Humanities*, en línea: https://www.dariah.eu (15/02/2024).

EDIT 16 = *Censimento nazionale delle edizioni italiane del XVI secolo*, en línea: https://edit16.iccu.sbn.it/web/edit-16 (15/02/2024).

ELEXIS = *European lexicographic infrastructure*, en línea: https://elex.is (15/02/2024).

GALLINA, Anna Maria (1959) *Contributi alla storia della lessicografia italo-spagnola dei secoli XVI e XVII*, Firenze, Olschki.

GILI GAYA, Samuel (1957) *Tesoro lexicográfico (1492-1726)*, Madrid, Consejo Superior de Investigaciones Científicas.

GÓMEZ ASENCIO, José Jesús (2007) "La edición de textos clásicos y su contribución al desarrollo de la historiografía lingüística", en Josefa Dorta, Cristóbal Corrales Zumbado y Dolores Corbella Díaz, eds., *Historiografía de la lingüística en el ámbito hispánico*, págs. 479-499, Madrid, Arco Libros.

ISASI VELASCO, Jennifer y Gimena DEL RIO RIANDE (2022) "¿En qué lengua citamos cuando escribimos sobre Humanidades Digitales?", *Revista De Humanidades Digitales 7*, págs. 127–143, en línea: https://revistas.uned.es/index.php/RHD/article/view/36280 (29/01/2024).

KHEMAKHEM, Mohamed et al. (2017) "Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields", *Electronic lexicography*, *eLex 2017*, Sep 2017, Leiden, Netherlands, en línea: https://hal.archives-ouvertes.fr/hal-01508868v2 (29/01/2024).

—— (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*, Austria: Graz, en línea: https://hal.archives-ouvertes.fr/hal-02263276 (29/01/2024).

LOPE BLANCH, Juan M. (1990) "El *Vocabulario de las dos lenguas toscana y castellana* de Cristóbal de las Casas", en Juan M. Lope Blanch, ed., *Estudios de historia lingüística hispánica,* Arco/Libros, Madrid, págs. 111-124.

LOMBARDINI, Hugo y Félix SAN VICENTE (2015) *Gramáticas de español para italófonos (siglos VI–XVIII). Catálogo crítico y estudio*, Münster, Nodus Publikationen.

NALESSO, Giulia (en prensa) "Humanidades digitales y lexicografía bilingüe: recuperación y valorización del patrimonio lexicográfico español-italiano (REVALSI)", en Alejandro Fajardo Aguirre, Dolores Torres Medina y Cristian Díaz Rodríguez, eds., *Lexicografía del español: internacionalización y contrastes*, Frankfurt am Main, Peter Lang.

—— (en prensa) "Obras lexicográficas para el aprendizaje del español en el siglo XVI: el caso del "Vocabulario de las dos lenguas" de Las Casas (1570)", en Giulia Nalesso y Alessandra Vicentini, eds., *Text and Ideas in the History of Language Teaching and Learning*, Bolonia, CLUEB.

NIEDEREHE, Hans-Josef (1994) *Bibliografía cronológica de la lingüística, la gramática y la lexicografía del español* / vol. 1, *(BICRES): Desde los principios hasta el año 1600*, Amsterdam/Philadelphia, John Benjamins.

DOI: http://dx.doi.org/10.13135/1594-378X/9988
Artifara, ISSN: 1594-378X

NIEDEREHE, Hans-Josef (1999) *Bibliografía cronológica de la lingüística, la gramática y la lexicografía del español* /vol. 2, *(BICRES II): Desde el año 1601 hasta el año 1700*, Amsterdam/Philadelphia, John Benjamins.

NIETO JIMÉNEZ, Lidio y Manuel ALVAR EZQUERRA (2007) *Nuevo tesoro lexicográfico del español (S. XIV.-1726)*. Madrid, Arco/Libros.

OVI = *Opera del Vocabolario Italiano* / Accademia della Crusca, en línea: http://www.ovi.cnr.it/Home.html (15/02/2024).

OXYGEN, en línea: https://www.oxygenxml.com (03/03/2024).

READ = *European Commission Horizon 2020 Research and Innovation Programme*, *Recognition and Enrichment of Archival Documents*, en línea: https://readcoop.eu (03/03/2024).

REAL ACADEMIA ESPAÑOLA, *Diccionario de la lengua española (DLE)*, 23.ª ed., en línea: https://dle.rae.es (10/03/2024).

——— *Nuevo Tesoro Lexicográfico de la Lengua Española (NTLLE)*, en línea: https://apps.rae.es/ntlle/SrvltGUISalirNtlle (03/03/2024).

REBIUN = Red de Bibliotecas Universitarias y Científicas Españolas, en línea: https://www.rebiun.org (10/03/2024).

SAN VICENTE, Félix (2010) "Diccionarios y didáctica en la tradición italoespañola (siglos XVI-XVII)", en Stefan Ruhstaller y María Dolores Gordon, eds., *Diccionario y aprendizaje del español*, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien, Peter Lang, págs. 47-88.

TEI = *Text Encoding Initiative*, en línea: https://tei-c.org (15/02/2024).

TLEAM = *Tesoro Lexicográfico del Español en América* / CORBELLA DÍAZ, Dolores Covadonga (dir.): https://portalciencia.ull.es/proyectos/51786/detalle (15/03/2024).

TRAMULLAS, Jesús (2002) "Propuestas de concepto y definición de la biblioteca digital.", *III Jornadas de Bibliotecas Digitales JBIDI*, San Lorenzo del Escorial, Universidad Politécnica de Madrid.

WILKINSON, Mark D. et al. (2016) "The FAIR guiding principles for scientific data management and stewardship", *Sci. Data3*, 160018, en línea: https://www.nature.com/articles/sdata201618 (15/02/2024).

WORLDCAT, en línea: https://www.worldcat.org/it (15/02/2024).