Revista de Humanidades Digitales

Del fonema al verso: una caja de herramientas digitales de escansión teatral[1]

From Phoneme to Verse: A Digital Scansion Toolbox for Plays

Fernando SANZ-LÁZARO
Universität Wien
fernando.sanz-lazaro@univie.ac.at
https://orcid.org/0000-0002-8815-6741

RESUMEN

Este artículo presenta los métodos y herramientas empleados en el tratamiento de textos teatrales en el proyecto Sound and Meaning in Spanish Golden Age Literature para disponerlos para el análisis digital. El proceso parte de textos obtenidos en archivos de muy diversos formatos y termina produciendo un archivo CSV apto para análisis estadísticos de lectura distante. Asume para ello del principio de dividir problemas complejos en la suma de tareas simples y abordar estas por separado. Se discute la limpieza y preparación de los archivos de origen según sus características propias. Se presentan las herramientas desarrolladas para abordar el análisis y clasificación de las unidades textuales automáticamente y el marco ontológico requerido para representar la estructura métrico-dramática. Finalmente, se detallan individualmente las bibliotecas envueltas en la escansión métrica, dado que la modularidad del proceso, que aborda aspectos muy específicos pero comunes a muchos ámbitos de los estudios lingüísticos, da pie a emplearlas por separado para propósitos distintos de para el que fueron concebidas.

PALABRAS CLAVE
Escansión automática, fonología, métrica, PLN, prosodia.

ABSTRACT

This article presents the methods and tools used to process theatrical texts by the project Sound and Meaning in Spanish Golden Age Literature to prepare them for digital analysis. The process starts from texts obtained in files of many different formats and finishes by producing a CSV file suitable for distant reading statistical analysis. For this, it assumes the principle of *divide et impera*, dividing complex problems into the sum of simple tasks and addressing them individually. The discussion addresses the cleaning and preparation of the source files according to their characteristics and the tools developed to automatically address the analysis and classification of textual units as well as the ontological framework required to represent the metric-dramatic structure. Finally, the libraries involved in metric scansion are presented individually, as the modularity of the process –which addresses very specific but common aspects of many fields related to linguistic studies– allows using them separately for purposes other than for which they were initially conceived together.

KEYWORDS
Automatic Scansion, Phonology, Metrics, NLP, Prosody.

RHD 8 (2023)

ISSN
2531-1786

[1] Publicado como parte de los proyectos *Sound and Meaning in Spanish Golden Age Literature* (FWF Austrian Science Fund, P 32563) y *The Interpretation of Childbirth in Early Modern Spain* (FWF Austrian Science Fund, P32263-G30).

Revista de Humanidades Digitales http://revistas.uned.es/index.php/RHD/index
Recibido 25/06/2023 – Aceptado 23/07/2023
Editado por G. del Rio Riande

1. INTRODUCCIÓN

El proyecto Sound and Meaning in Spanish Golden Age Literature se ocupa de obras teatrales del Siglo de Oro (Kroll y Sanz-Lázaro, 2022; 2023). Aborda el texto dramático considerando también las características métricas de sus versos. Entre otras técnicas, se aproxima al objeto de estudio mediante la lectura distante (Moretti, 2000, pp. 57-58). Para ello, necesita disponer de un corpus adecuado sobre el que llevar a cabo los análisis. Sin embargo, no había un corpus de tales características disponible al comienzo de las investigaciones, por lo que se hizo necesario producir uno. Recopilar un corpus es, no obstante, uno de los mayores retos que afrontan los análisis digitales por la inversión de recursos humanos y temporales que requiere (Ehrlicher, 2019, p. 41). Dadas las características del proyecto y sus demandas particulares, se optó por llevar a cabo internamente el proceso en su totalidad, desde la recopilación de obras teatrales, pasando por su preparación, estructuración y clasificación hasta disponer de datos aptos para llevar a cabo los análisis proyectados.

Se optó por emplear de manera exclusiva textos procedentes de ediciones críticas para garantizar la mayor calidad posible de las fuentes. Esto agiliza el trabajo, pues permite asumir que la edición es óptima, por lo que el texto puede emplearse tal como está sin necesidad de correcciones previas. Como contrapartida, el corpus resulta más reducido como consecuencia de una mayor lentitud de crecimiento.

El primer reto que se plantea es preparar los textos digitalizados. La extracción y copia manual de cientos de miles de versos es un trabajo largo y tedioso. De ahí que surja la necesidad de automatizar las labores envueltas en la tarea, de manera que el investigador tenga que dedicarle el menor tiempo posible y puede concentrarse en interpretar los resultados. Sin embargo, no es preciso tratar el proceso como si fuera un ente monolítico, sino que puede descomponerse en diferentes subprocesos para resolver cada uno de ellos con herramientas específicas para el trabajo: *divide et impera*. Se pueden emplear aplicaciones preexistentes, si las hay, y desarrollarlas, en caso contrario.

El objetivo es crear pequeñas herramientas informáticas orientadas a cometidos muy específicos con el fin de combinar sus resultados parciales para alcanzar el final. Esto presenta varias ventajas: entre otras, el desarrollo se agiliza, pues simplifica la detección y corrección de fallos. Sin embargo, para el potencial usuario, resulta también positivo, pues las microtareas pueden ser partes constitutivas de otras labores diferentes o incluso ser trabajos finalistas por sí mismas. Así, los módulos pueden usarse por separado para fines diversos, mientras que, de organizarse todo en un programa monolítico, solo serviría para resolver los problemas particulares para los que fueron diseñadas. Piénsese en un juego de construcciones: una construcción terminada y sin posibilidad de modificarla no puede ser más que eso, mientras que, si se dispone de todos los bloques con los que se construyó, puede hacerse la misma construcción o una infinidad de otras figuras distintas. De ahí el peso que recibe la estructura modular para multiplicar los usos posibles.

El principio rector es descomponer un problema complejo en varios simples y emplear instrumentos sencillos para resolverlos por separado.

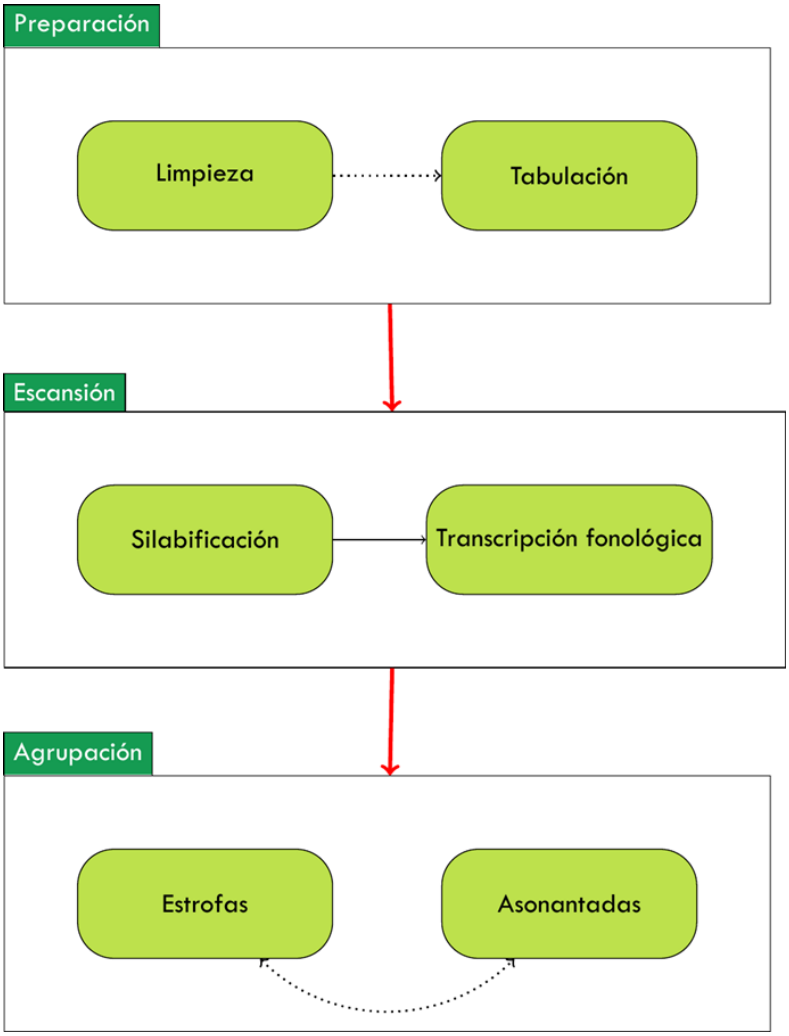


Figura 1. Cadena de trabajo. Fuente: elaboración propia.

Esto no quiere decir que haya que dar todos los pasos manualmente, sino que su ejecución y la comunicación entre ellos –dirigir los datos de salida de una etapa a la entrada de la siguiente– puede automatizarse. Por este motivo, en lugar de programas ejecutables, resulta más conveniente crear bibliotecas de funciones listas para ser ensambladas y usadas por otros programas. No se asume el uso que se les va a dar, sino que se ofrecen tal cual para ser empleadas de la manera que mejor convenga en cada momento y para cada propósito. Para facilitar la manipulación de los textos, las bibliotecas tienen un diseño orientado a objetos, con clases para caracterizar de forma intuitiva las entidades textuales o lingüísticas y las operaciones que se hacen sobre ellas. Todas ellas se han publicado bajo diferentes variantes de licencia GNU GPL y LGPL, lo que garantiza la libre distribución de los programas y la disponibilidad de su código fuente. Este, así como instrucciones detalladas de uso, es accesible en línea en los repositorios GitHub de cada una de las piezas de software.

[2] Se trata de dos licencias que garantizan que el código es libre y abierto. En concreto, la licencia GPL permite usar, estudiar, compartir y modificar el programa, pero obliga a distribuir los trabajos derivados bajo los mismos términos o equivalentes. La LGPL es menos restrictiva, pues permite integrar el software en forma de biblioteca externa en programas con cualquier tipo de licencia, incluso comercial, sin necesidad de que adopte también una licencia libre o publique su código, salvo las partes cubiertas por la LGPL, pues ha de respetar el carácter libre y abierto de estas (Free Software Foundation, 2022).

2. PREPARACIÓN PRELIMINAR DEL TEXTO

¿Cómo se desarrolla este trabajo? A grandes rasgos, el proceso a seguir es el siguiente: se toma el texto digitalizado de una obra teatral y se limpia para disponerlo en un formato conveniente para su manipulación; se analiza la forma externa de la obra buscando entidades estructurales para tabularlos; se analizan los versos de la tabla y se añaden columnas con la información métrica de los versos; finalmente, se coteja esa información para agrupar los versos en unidades mayores, como estrofas o poemas (Figura 1). Aunque la estrategia resulta sencilla, incluso previsible, la forma de llevarla a la práctica no lo es tanto. Para empezar, las fuentes suelen presentarse en los formatos más diversos, desde XML-TEI, en el mejor de los casos, hasta archivos PDF de difícil manejo, pasando por casos tan peculiares como documentos para WordPerfect 5.2, un paquete ofimático que salió al mercado en 1992.

2.1. Peor escenario: PDF

Los archivos PDF codifican el texto en un formato binario que no responde a la estructura textual, sino a aspectos visuales. Requieren un programa para interpretar el código del archivo y representarlo en la pantalla. No es posible trasladar la composición de la página de forma directa, por lo que hay que convertir la fuente a texto plano y hacerlo de tal manera que la contribución manual se reduzca tanto como sea posible. Para ello, hay que entender cómo trabaja el formato internamente: el archivo contiene una serie de instrucciones que definen los bloques visuales representados en pantalla. En el caso del texto, estos bloques son segmentos de caracteres, pero no se ordenan de acuerdo a la estructura textual, sino la de maquetación. La complejidad de la composición de la página y su fidelidad a la jerarquía textual determinan la forma de abordar el texto.

Si hay un grado de equivalencia grande entre la estructura textual y la visual, valdría emplear las herramientas habituales para convertir entre formatos. Después, como es natural, habría que afinar el trabajo a mano, aunque, por lo general, es posible hacerlo mediante sustituciones con expresiones regulares. Hay que eliminar la paginación, las cabeceras de las páginas, numeración de los versos, aparato crítico, números volados de llamada a nota, etc., pero conservando los parlamentos y didascalias. Resulta aquí de gran ayuda si, como resultado de la conversión de formato, los saltos de página vienen ya marcados mediante un meta carácter. De ser así, es posible orientarse por este y los interlineados para identificar cabeceras y notas.

El cuerpo del texto, por su parte, está constituido por varias entidades diferenciadas que hay que identificar por separado. Las líneas y las didascalias de personaje se reconocen por el espaciado: la didascalia se indica mediante texto sin sangrar y la línea de diálogo con sangrado, salvo la primera del parlamento, que está varios espacios tras la didascalia que la introduce. En caso de versos compartidos, el programa empleado y la codificación del archivo de origen son decisivos para evitar la intervención manual. Si la herramienta de conversión interpreta el

espaciado de origen como mayor que el normal, este aparecerá en el archivo resultante. De no ser así, faltará la distinción entre un verso normal y la continuación de uno compartido, por lo que será necesario identificarlo visualmente para señalarlo a mano.

Las acotaciones escénicas requieren también una atención especial. Se distinguen de manera sencilla si van centradas en la página. Esto sucede a menudo con las iniciales, que introducen las circunstancias al comienzo de la obra o una jornada. Sin embargo, son menos comunes una vez la acción ha dado comienzo. Si este es el caso, el tratamiento se complica, pues el texto plano no contempla atributos de formato para los caracteres. Por consiguiente, no queda otro remedio que recurrir a la intervención manual en tal situación. Por fortuna, la mayoría de las acotaciones suelen ser reconocibles mediante una inspección visual superficial, ya que suelen diferir en tamaño de los versos colindantes, bien por ser significativamente más breves (*vase*, *canta*, *dentro*) o más extensas (descripciones de los elementos escénicos).

Conviene hacer la prevención de que este primer paso no tiene por qué darse necesariamente tomando los textos uno a uno. Bien al contrario, trabajar con lotes de archivos de la misma procedencia agiliza el trabajo, pues estos suelen haber sido compuestos con las mismas herramientas y criterios de maquetación, por lo que suelen presentar características similares. En una situación ideal, se determinarían de forma manual los patrones de búsqueda requeridos para hacer las sustituciones en un texto piloto, pero luego se aplicarían en masa al resto de textos análogos mediante guiones o *scripts*.

2.2. Texto y Visually Encoded Drama (VED)

En el mejor de los casos, el documento de origen se encuentra en alguna forma de archivo de texto, con o sin formato. Este tipo de datos ya presentan entidades textuales reconocibles: aunque sea mínimamente gracias a los saltos de línea, las entidades estructurales son textuales y no espaciales, como era en los PDF. Esto permite empezar a hacer sustituciones de texto para limpiar el documento y adaptarlo al siguiente paso de la serie. Si se halla en codificado como texto puro, el límite lo pone la claridad de la disposición de los elementos, en tanto que hay que identificarlos mediante expresiones regulares. En el caso de texto enriquecido de algún tipo, como un documento para procesador de texto, puede aprovecharse que algunos programas ofimáticos permiten combinar las expresiones regulares con atributos de formato, de manera que la labor de encontrar, por ejemplo, acotaciones –normalmente en cursiva– se simplificaría aún más.

En cualquier caso, al estar trabajando con sustituciones, la mayor complicación que se presenta es definir un patrón de búsqueda que corresponda a la cadena a encontrar, pero la sustitución es después discrecional. Por lo tanto, es factible introducir en la cadena sustituta marcas para diferenciar entre sí los componentes del texto. Surge así la posibilidad de un formato que no solo muestre el texto limpio, sino que aporte información adicional. Mediante una adición mínima, pueden categorizarse los componentes del texto y su metainformación.

Se trata, pues, de, una vez identificadas entidades cuya categorización y cuantificación

permita extraer información *dramétrica*[3] del texto para hacer análisis tanto de de lectura distante como estructurales de la obra, marcarlas de alguna manera para no perder esa información obtenida fortuitamente y aprovecharlo después. Esto es, se conservará la información que sirva como base para la búsqueda heurística de las diferencias en un corpus de textos o para encontrar patrones en una obra de forma individual (Ilsemann, 1998, pp. 15-16).

De esta manera, las reglas del formato VED (Visually Encoded Drama) no pueden ser más sencillas. A cada entidad textual (línea, didascalia de personaje, otro tipo de acotación, jornada, etc.) o metatextual (autor, título, fecha, etc.) le corresponde una única línea. Los textos declamados van sangrados (una vez para comienzo de verso, dos o más en las continuaciones) y las didascalias de personaje no. Adicionalmente, hay etiquetas para indicar acotaciones de otro tipo, líneas no versificadas o fuera de la cuenta, como parlamentos en prosa o ecos, y elementos metatextuales. De esta manera, no solo se diferencian estructuralmente los elementos de la obra, sino que lo hacen de una manera que facilita sobremanera la inspección visual (Figura 2).

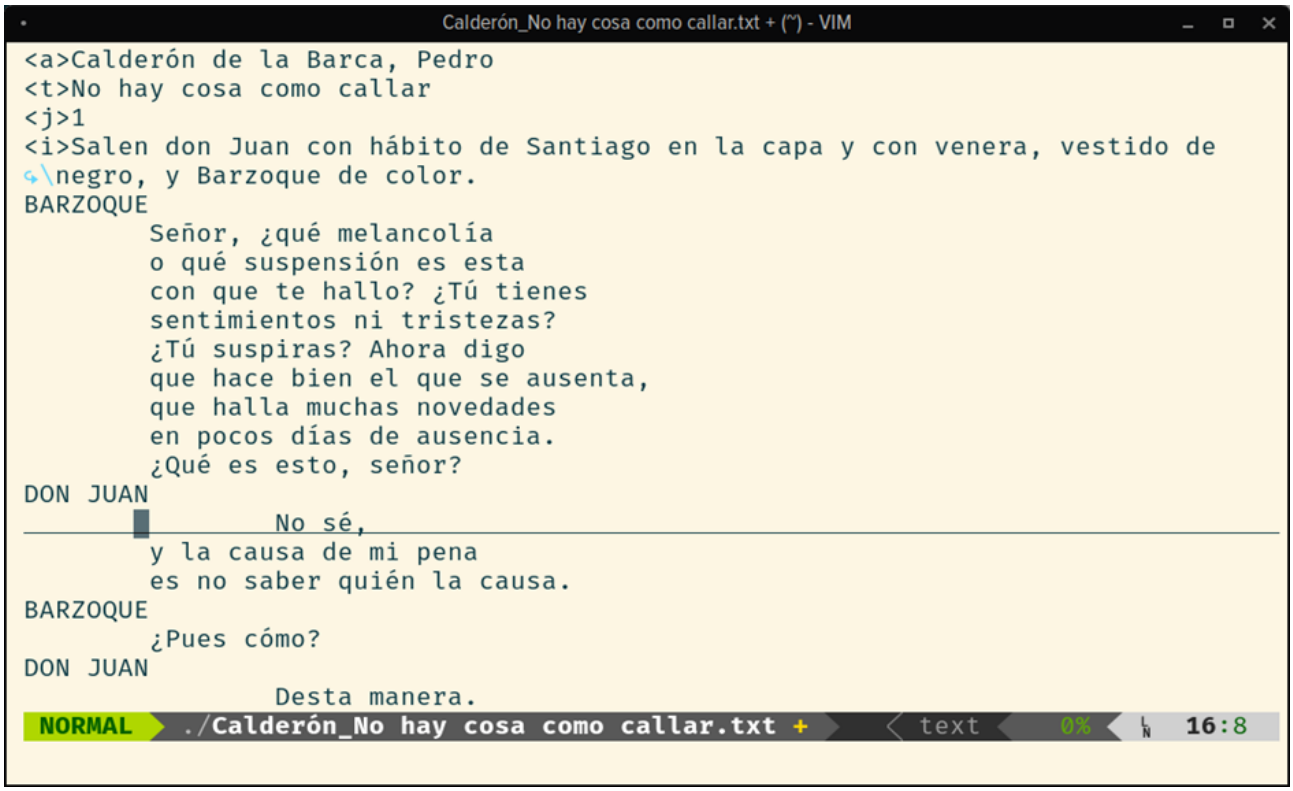


Figura 2. Archivo VED. Fuente: elaboración propia.

Las ventajas de este pequeño rodeo son múltiples. Primero, se hace durante la preparación de los textos sin que suponga un incremento en la carga de trabajo requerida. Han de tenerse en cuenta los recursos que demandaría convertir el texto de origen después de limpiarlo a, por ejemplo, XML-TEI. Segundo, no requiere un conocimiento especial, pues se basa en tabulaciones y un reducido número de etiquetas, la mayoría opcionales. Tercero, se presta a la inspección ocular rápida: el formato visual permite identificar líneas incoherentes y corregirlas rápidamente. Cuarto, es un formato semiestructurado apto para el tratamiento digital, por lo que permite la conversión automática a un formato tabular o a XML-TEI, mucho más apropiados tanto para el análisis como pa-

[3] Romanska (2015, p. 446) acuñó el término *dramétrica* para referirse a la aproximación matemática al análisis del drama.

ra la difusión del texto preparado. Valga como ejemplo la contribución de textos de Calderón de la Barca a la colección DraCor (Fischer et al., 2019), codificados mediante esta herramienta.

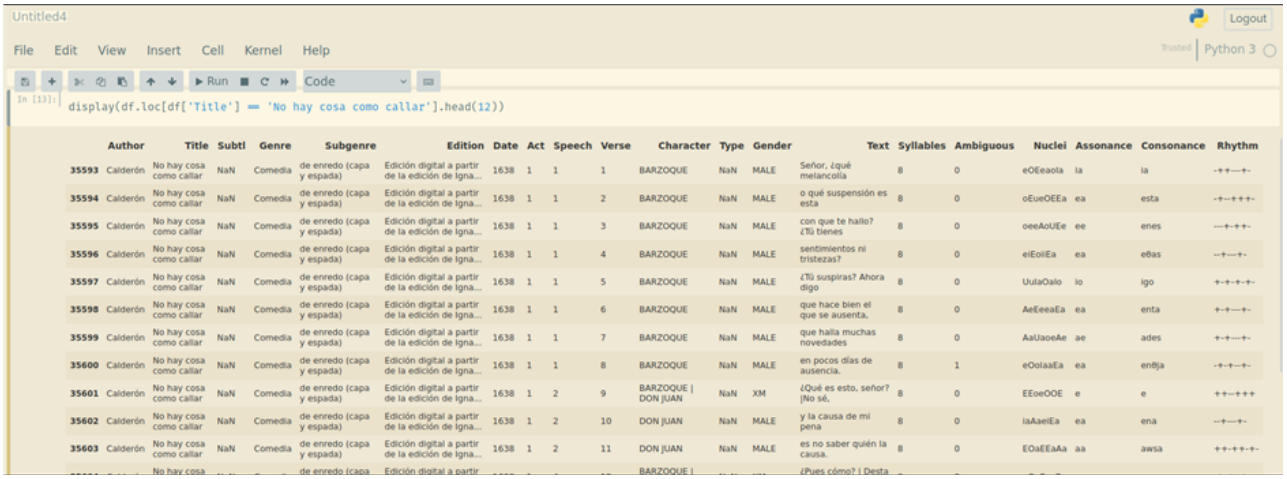
Aquí se da un paso intermedio con uso finalista fuera del contexto para el que se diseñó. El formato VED requiere un esfuerzo adicional mínimo para limpiar el texto, pero, a cambio de los pocos minutos adicionales requeridos para añadir algunas líneas con metadatos, como el autor o el título, permite codificar la obra teatral de manera automática en XML-TEI sin conocimientos específicos, ni para el texto de origen, que es intuitivo y no lo requiere, ni para el documento final, que se produce sin concurso humano. En consideración a esto, se ha publicado bajo licencia GPL el paquete PyPi (Python Software Foundation, 2023) *xml2tei* (Sanz-Lázaro, 2023c), que incluye tanto la biblioteca como un *script* ejecutable para crear archivos XML-TEI a partir de VED.

2.3. Caso óptimo: XML-TEI

Si, por fortuna, el texto ya se encuentra codificado como XML-TEI, la solución se antoja sencilla. El etiquetado –cuando es correcto y relativamente detallado– define claramente las entidades textuales necesarias para poder trabajar. En el caso del proyecto Sound and Meaning, estas son, como mínimo, el parlamento, y atributos de este, como el locutor, o subestructuras, como sus líneas de texto. Si se ofrecen otras entidades estructurales, como el *acto*, o metatextuales, como el *título* o *información editorial*, se alivia el trabajo. Su ausencia no supone un impedimento mayor, ya que una sola instancia por obra hace factible su inclusión manual. Por el contrario, tratar a mano todos los elementos estructurales de más bajo nivel, como el *parlamento* o la *línea*, supondrían una carga de trabajo inasumible en un corpus de cierto tamaño.

Un último elemento de gran impacto en el tiempo de proceso es el atributo *part* de la etiqueta TEI <l>. Este aparece en fragmentos de versos compartidos repartidos a lo largo de dos o más líneas. Adopta un valor diferente según la línea referida constituya el inicio, el final o el interior del verso. De esta manera, se distingue entre verso y línea como dos entidades diferenciadas. Si bien verso equivale a línea en ediciones históricas y manuscritos, la convención moderna es separar los parlamentos y sangrar las líneas que continúan un verso anterior. Dado que el trabajo se lleva a cabo sobre ediciones modernas, se hace imprescindible tener esto en cuenta.

El primer paso consiste pues en trasladar la estructura jerárquica de XML a forma tabular. La codificación XML-TEI representa la estructura textual sin ambigüedades. Esto es suficiente, ya que el propósito del trabajo no demandaba definir aspectos no lineales, como sí ocurre a menudo en otros ámbitos, como la ecdótica (Presotto, 2023, pp. 7-10). La estrategia es tomar la línea como unidad mínima y definir otros datos, estructurales o no, como sus atributos. De esta manera, cada línea de diálogo presenta características comunes a toda la obra, como el autor o el título, que la diferencian de líneas de otras obras. La estructura externa se denota mediante números enteros para ubicar la línea en el texto: en qué acto está, en qué parlamento o qué línea es. Otra columna indica características de la línea, como el locutor, cuyo valor comparten todas las líneas de un parlamento, y otra el número de verso.



	Author	Title	Subtl	Genre	Subgenre	Edition	Date	Act	Speech	Verse	Character	Type	Gender	Text	Syllables	Ambiguous	Nuclei	Assonance	Consonance	Rhythm
35593	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	1	1	BARZOQUE	NaN	MALE	Señor, ¿qué melancolía	8	0	eOEeaola	ia	ia	-+•--+-
35594	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	1	2	BARZOQUE	NaN	MALE	o qué suspensión es esta	8	0	oEueOEEa	ea	esta	-+•--+++-
35595	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	1	3	BARZOQUE	NaN	MALE	con que te hallo? ¿Tú tienes	8	0	oeeAoUEe	ee	enes	---+-++-
35596	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	1	4	BARZOQUE	NaN	MALE	sentimientos ni tristezas?	8	0	eIEoIEa	ea	eßas	--+---+-
35597	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	1	5	BARZOQUE	NaN	MALE	¿Tú suspiras? Ahora digo	8	0	UuIaOaIo	io	igo	+-+-+-+-
35598	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	1	6	BARZOQUE	NaN	MALE	que hace bien el que se ausenta,	8	0	AeEeeaEa	ea	enta	+-+---+-
35599	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	1	7	BARZOQUE	NaN	MALE	que halla muchas novedades	8	0	AaUaoeAe	ae	ades	+-+---+-
35600	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	1	8	BARZOQUE	NaN	MALE	en pocos días de ausencia.	8	1	eOoIaaEa	ea	enθja	-+•--+-
35601	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	2	9	BARZOQUE | DON JUAN	NaN	XM	¿Qué es esto, señor? ¡No sé,	8	0	EEoeOOE	e	e	++--+++
35602	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	2	10	DON JUAN	NaN	MALE	y la causa de mi pena	8	0	IaAaeIEa	ea	ena	--+---+-
35603	Calderón	No hay cosa como callar	NaN	Comedia	de enredo (capa y espada)	Edición digital a partir de la edición de Igna...	1638	1	2	11	DON JUAN	NaN	MALE	es no saber quién la causa.	8	0	EOaEEaAa	aa	awsa	++-++-+-
		No hay cosa				Edición digital a partir					BARZOQUE |			¿Pues cómo? ¡ Desta						

Figura 3. Obra en formato tabular. Fuente: elaboración propia.

El resultado es una tabla con la información requerida (Figura 3). La tabla contiene columnas con atributos invariables de la comedia, como el autor o el título; información estructural, como el número de jornada, de parlamento o de verso, y datos dramáticos, como el texto, el locutor que lo pronuncia o las características de este. A esto se añadirá en la siguiente fase la información métrica correspondiente a cada verso.

3. ANÁLISIS MÉTRICO

Queda la cuestión métrica. Disponiendo de la tabla, los datos necesarios se encuentran listos para ser analizados según los principios que bosquejamos someramente en otro lado (Sanz-Lázaro, 2023b). Solo hay que recorrerla y, en caso de encontrar filas distintas con número de verso coincidente para la misma obra, tomar el texto de todas ellas como un único verso, pues se trataría de uno compartido. Sin embargo, el drama presenta particularidades que no se dan en el verso lírico, como la polimetría, y tendremos que hacernos cargo de ellos.

La escansión automática de poesía española, desde sus primeros intentos serios (Gervás, 2000), ha visto un desarrollo meteórico hasta alcanzar su madurez metodológica en menos de veinte años (Navarro-Colorado, 2015; 2017), cuyas variantes siguen rindiendo los mejores resultados (Rosa et al., 2020). No obstante, se han abierto nuevas vías de investigación (Agirrezabal, Alegría y Hulden, 2017; Marco Remón y Gonzalo Arroyo, 2021; Rosa et al., en prensa) que, sin llegar a alcanzar la precisión del modelo clásico (Marco Remón et al., 2021), dejan entrever que no todo está dicho y aconsejan profundizar en ellas y seguir experimentando con otras nuevas.

La aproximación aquí descrita presenta una reinterpretación del modelo clásico iniciado por Navarro-Colorado –por ser este el que mejores resultados sigue ofreciendo por el momento– que aborda las particularidades del teatro, así como otras generales de la prosodia española. Dicho de otro modo, adopta la silabación para determinar las unidades rítmicas y la morfosintaxis para caracterizarlas. Si bien nuestra aproximación contempla esa organización abstracta del proceso, la concreta mediante procedimientos novedosos en una biblioteca propia para escandir versos. Para simplificar el trabajo, lo hace desde una novedosa aproximación fonológica en lugar de la grafémica usada hasta ahora. El lector humano, al escandir un verso escrito, no prescinde de la fonología, sino que la ofusca en los procesos mentales bajo subterfugios ortográficos. Al ponderar

una sinalefa, se traducen los grafemas del verso a su expresión fónica, que cada cual evalúa según su conciencia lingüística para determinar la factibilidad de resoluciones potenciales. Ante vocales en contacto, por ejemplo, estas se desilabizan una a una para diptongar la unión. El tratamiento digital obligaría a formalizar esto y, además, a traducir la evaluación fonológica implícita en una representación grafémica. Siendo así, resulta mucho más sencillo trabajar explícitamente con transcripciones fonológicas que disfrazar la oralidad con un manto grafémico. En palabras de la germanista y premiada traductora literaria Isabel García Adánez (2005): "en la poesía, la materia prima es la sustancia fonética, el plano de la articulación" (p. 96).

3.1. Transcripción y silabación

Para dividir palabras en sus sílabas, hemos creado la biblioteca *Silabeador* (Sanz-Lázaro, 2022b)[4]. Como su propio nombre indica, separa palabras en sílabas. ¿Por qué surge la necesidad de diseñar un nuevo programa para esta función? Volviendo a la fonología, los programas disponibles se basan en la ortografía y, de ahí, que no tengan en cuenta la realización real de las palabras sino un sistema de representación incompleto (RAE y ASALE, 2010a, § 4.4.2b). En su día, las Academias acordaron unas reglas ortográficas simples y universales. Si bien esto simplifica y unifica la escritura, la realidad demuestra ser más complicada en el caso de la vertiente hablada, pues dichas reglas no cubren todos los casos. Ante la tesitura de tener que reconocer excepciones para representar el objeto fielmente, se optó por lo contrario: tratar de amoldar el objeto de estudio al modelo. De ahí la necesidad de figuras como los llamados *diptongos ortográficos* (RAE y ASALE, 2010b, § 3.4.2.2.1). A despecho del elegante modelo ortográfico de la Real Academia, la declamación ideal de los versos responde en esos casos a las convenciones de la lengua y no al del sistema de representación académico. Por lo tanto, si el modelo no se adecúa al propósito, resulta sensato cambiar la representación antes que pronunciar el verso en falso. Este es el motivo de que la biblioteca de división silábica aplique una combinación de reglas y diccionario para dividir donde es necesario. Esto permite además reconocer algunos latinismos sin adaptar y tratar las palabras, de acuerdo a su sistema fonológico, tanto para los diptongos como para la acentuación cuando la cantidad silábica se deduce de la ortografía.

Por otra parte, algunos de los signos ortográficos más exóticos suelen desencadenar errores en la silabación. En el mejor de los casos, el programa es incapaz de completarse al fallar la lectura de la cadena de caracteres. Afortunadamente, esto permite identificar la causa y tratarla en consecuencia. En el peor de los casos, el programa termina su ejecución sin interrupciones, pero devuelve una lista de sílabas incorrecta. Corregir esta última situación presenta grandes dificultades, pues las palabras individuales suelen pasar desapercibidas en un corpus de cierto volumen. Sin embargo, en los textos teatrales, aparecen desde extranjerismos sin adaptar hasta diacríticos específicos de la poesía, como cremas para denotar diéresis. Por lo tanto, el procesamiento adecuado del drama áureo exige considerar esos casos de manera adecuada. La biblioteca *Silabeador*

[4] Accesible desde: https://github.com/fsanzl/silabeador.

ofrece una precisión del 99,92 % al analizar el corpus EDFU (Rosa y Pérez, 2020) de más de 100.000 palabras. No solo eso, sino que el porcentaje de errores se debe exclusivamente a palabras extranjeras sin adaptar (principalmente de lenguas germánicas como el alemán o el inglés).

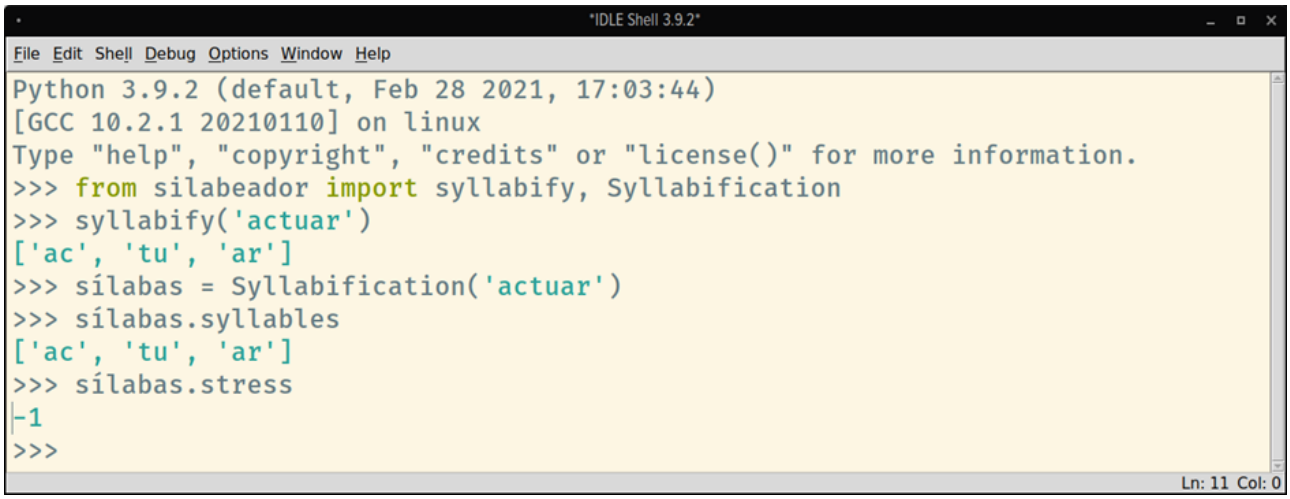
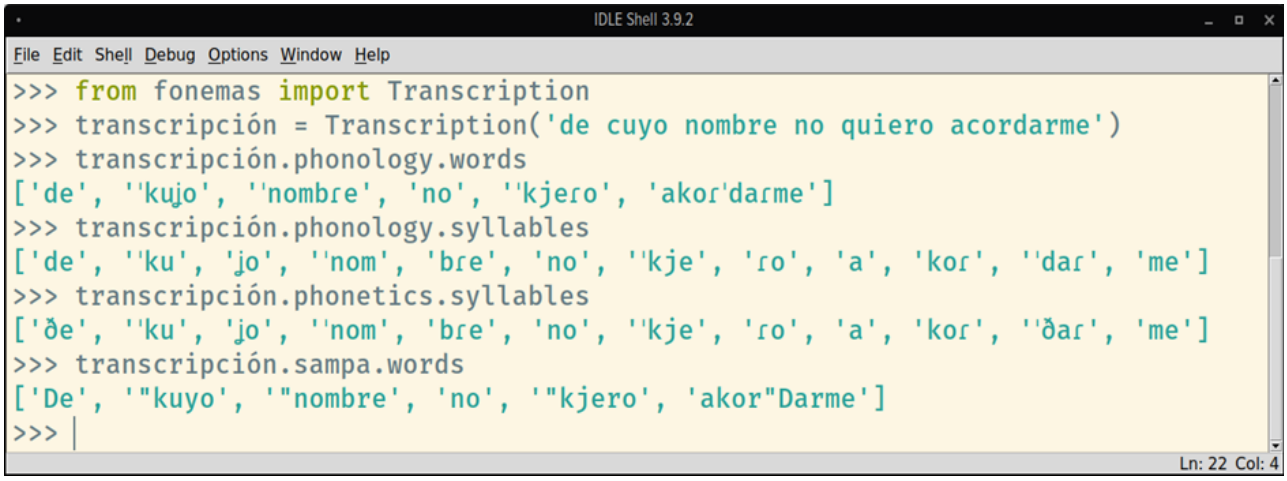


Figura 4. Ejemplo de silabación. Fuente: elaboración propia.

El modo de funcionamiento es simple. Se instala la biblioteca a partir de su código fuente o como un paquete PyPi, se importa en el programa y se crea una instancia de la clase *Syllabification*, pasándole como argumento la palabra a analizar. Esto crea un objeto con dos atributos, *syllables* y *stress*, que son, respectivamente, una lista de sílabas y un número entero negativo, que adopta valores entre -1 y -3, según si ocupa la última, penúltima o antepenúltima sílaba (Figura 4). La clase admite argumentos facultativos para modificar su comportamiento, como la forma de agrupar el grupo *tl* –posible *en ataque* en Centroamérica–, la silabación y acentuación de, por ejemplo, *scriptum*, según las reglas ortográficas españolas o latinas, y el tratamiento del grupo *sch* alemán *en ataque*.

La biblioteca anterior sirve como base a otra, llamada *Fonemas* (Sanz-Lázaro, 2022b). Su nombre es también descriptivo de su función. En pocas palabras, se trata de un módulo de transcripción fonológica. Presenta la ventaja de que explota el estándar Unicode para presentar los resultados en el Alfabeto Fonético Internacional, si bien con algunas peculiaridades por necesidades prácticas. Los sistemas operativos modernos han superado las restricciones que imponían las codificaciones de caracteres de antaño, por lo que tiene poco sentido autolimitarse en este sentido. No obstante, para garantizar la retrocompatibilidad, el módulo produce también la transcripción SAMPA, que puede resultar útil para algunas aplicaciones informáticas. Además, incluye algunos extras, como la transcripción fonética (parcial por ahora). La transcripción se toma algunas libertades respecto al AFI, como marcar /i̯/ y /u̯/ semivocálicas con el mismo símbolo que las semiconsonantes /j/ y /w/ correspondientes. Esto responde a una razón práctica: el diacrítico Unicode del sonido semivocálico *combining inverted breve below* requiere un carácter normal y un combinado, lo que, en realidad, son al final dos caracteres. Si el programa que toma la transcripción no se lo espera, interpretaría una secuencia de dos caracteres, lo que podría tener consecuencias impredecibles. Por la misma razón, representamos el resto de vocoides mediante el diacrítico de breve (arriba), ya que existen caracteres simples para ello, mientras que se requie-



ren dos para ponerles el diacrítico abajo.

Figura 5. Ejemplo de transcripción. Fuente: elaboración propia.

La biblioteca se usa de forma análoga a la anterior. Después de instalar la biblioteca como paquete PyPi o descargando su código fuente, se importa la clase *Transcription* y se crea una instancia pasándole una palabra o, ahora también, una frase. El objeto tiene tres atributos *phonology*, *phonetics* y *sampa*, cada uno de ellos con los subatributos de nombre autodescriptivo *words* y *syllables* (Figura 5).

3.2. Primera pasada

La transcripción fonológica es solo uno de los fundamentos de otras operaciones más complejas. En concreto, aquí se emplea como base para la escansión métrica de los versos. Esta se lleva a cabo mediante la biblioteca *libEscansión* (Sanz-Lázaro, 2023a)[5], que, a su vez, delega la transcripción del texto de entrada en la antes descrita. La biblioteca hace una escansión *natural* del verso según el modelo silábico-gramatical. Este modelo, por otra parte, resulta una manera factible de acercarnos a la métrica desde el texto impreso, como una idealización de la realización sonora física –mejor dicho, cada una de sus realizaciones–, que solventa las dificultades de esta, si bien a costa de generalizar (Sánchez Jiménez, 2017). Esto es, divide las palabras en sus sílabas y determina el acento interno, que deviene en métrico o no, de acuerdo a la categoría gramatical de la palabra en cuestión. La consideración superficial de las relaciones sintácticas supone una mejora en compuestos de la misma categoría y la aproximación fonológica simplifica la resolución de ambigüedades. Los resultados obtenidos superaron con creces las expectativas. Se puso a prueba la precisión usando el corpus ADSO 100 (Navarro-Colorado, Ribes Lafoz y Sánchez, 2016), compuesto de cien sonetos escandidos con un 96%, de acuerdo a anotadores humanos. La biblioteca alcanzó una precisión del 97,01% con la configuración estándar y 98,65% forzando la interpretación átona de las interjecciones.

Puede ilustrarse de forma sucinta mediante la reducción de vocales en contacto entre palabras según el sistema de evaluación de uniones potenciales: en ese caso, busca primero las sinalefas y sinéresis posibles en el verso considerando la escala de perceptibilidad. A cada unión poten-

[5] Accesible desde: https://github.com/fsanzl/libEscansion.

cial se le asigna un índice de precedencia. Este índice depende principalmente de la acentuación de cada una de las vocales en contacto, del número de fonemas que integren el clúster vocálico resultante y de la distancia en el trapecio vocálico de las vocales liminares. Este sistema es dinámico: no toma todas las sinalefas encontradas en el verso y las resuelve según un criterio dado hasta cuadrarlo, sino que resuelve la unión de prioridad más alta y evalúa de nuevo el verso resultante, asignando prioridades que tienen en cuenta la nueva unión. Los versos resultantes de cada corrección se siguen evaluando recursivamente hasta llegar al metro buscado.

La forma de usar la biblioteca *libEscansión* no difiere de lo visto para las anteriores. Después de instalarla, se importa la clase *VerseMetre* y se crea una instancia de ella que espera recibir un verso como argumento. El objeto resultante tiene como atributos los elementos caracterizadores del verso. Asimismo, la clase admite parámetros para modificar el comportamiento, como, por ejemplo, forzar un número determinado de sílabas (Figura 6).

```
File Edit Shell Debug Options Window Help
>>> from libEscansion import VerseMetre
>>> verso = VerseMetre('alma a quien todo un dios prisión ha sido')
>>> verso.syllables
[['Al', 'ma'], ['kjen'], ['tO', 'dŏUn'], ['djOs'], ['pɾi', 'sjOn'], ['ʰA'], ['sI'
, 'do']]
>>> verso.rhythm
'+--+++-+++-'
>>> verso.nuclei
'AaeOUOiOAIo'
>>> verso.rhyme
'ido'
>>> verso.asson
'io'
>>> verso.synaloephas
[([0, 1], 4.0, 'ma', 'a'), ([3, 1], -4.0, 'do', 'Un')]
>>> verso.count
11
>>> verso.line
'alma a quien todo un dios prisión ha sido'
>>> verso = VerseMetre('alma a quien todo un dios prisión ha sido', [14])
>>> verso.syllables
[['Al', 'ma'], ['a'], ['kjen'], ['tO', 'do'], ['Un'], ['djOs'], ['pɾi', 'si', 'O
n'], ['ʰA'], ['sI', 'do']]
>>>
                                                                    Ln: 25 Col: 4
```

Figura 6. Ejemplo de escansión de un verso. Fuente: elaboración propia.

El rendimiento con metro variable –no solo poemas heterométricos, que analiza con una precisión equivalente a la de los poemas isométricos, sino con textos polimétricos, como las obras teatrales– es excelente. A diferencia de otros métodos de escandir, *libEscansión* no se limita a aplicar reglas de la poética española ni intenta satisfacer un metro fijo. Por el contrario, combina ambas aproximaciones aplicando las reglas para intentar resolver de acuerdo a una serie de expectativas ordenadas según su preferencia. Emplea una lista ordenada de metros esperables que intenta armonizar con los ajustes potenciales, de manera que, si no es posible resolver con un metro dado, lo intenta nuevamente desde el principio con el siguiente provisto en la lista. Esto permite también considerar primero los metros habituales, como octosílabos o endecasílabos, y solo intentará resoluciones más exóticas (eneasílabos, por ejemplo) de no poder hacer las situadas primero.

La flexibilidad de este mecanismo permite utilizar un tercer programa, tal y como se usa en el proyecto Sound and Meaning, que carga la biblioteca de escansión y va ajustando al vuelo los parámetros para cada nuevo verso, de acuerdo al contexto métrico y dramático. De esta manera, si el locutor del parlamento o las acotaciones sugieren una parte cantada, se intentará resolver el metro primero como hexasílabo –que es habitual en este tipo de versos–o, si aparece un verso endecasílabo, se modificará la lista de versos, de manera que no solo suba *11* a la primera posición, sino que la segunda la ocupará *7*, ya que endecasílabos y heptasílabos se agrupan con frecuencia, en, por ejemplo, silvas.

3.3. Segunda pasada

La segunda pasada tiene dos objetivos. Por un lado, encuentra agrupaciones de versos. Por otro, corrige posibles errores de la pasada previa. Para lo primero, evalúa los versos y las rimas consonantes para identificar esquemas de rima consonante en números dados de versos que denoten una estrofa. Empieza intentando localizar las estrofas de más versos y continúa con las de menos de manera sucesiva. Los versos que no han sido marcados se evalúan de nuevo en busca de romances y otras formas arromanzadas. Con los versos agrupados, pueden cotejarse los versos libres de ambigüedades y tomar su metro como medida para escandir nuevamente otros versos de la agrupación con un metro incoherente con los versos adyacentes. Por ejemplo, dado que la escansión considera el entorno, es posible que el primer verso de una silva tras un romance haya sido resuelto como octosílabo y, viceversa, que el primer verso de un romance tras una silva haya sido evaluado como heptasílabo. Sabiendo por la rima a cuál de ellos pertenecen, puede evaluarse nuevamente el octosílabo inicial de la silva como heptasílabo o el heptasílabo inicial del romance como octosílabo.

4. CONCLUSIONES

Hemos presentado métodos y herramientas que, puestos en combinación, permiten llevar a cabo la composición de un corpus dramático anotado del Siglo de Oro mediante el análisis estructural y métrico de obras teatrales. El proceso consta de varios pasos, cada uno de los cuales prepara los datos de entrada para el siguiente. Primero, se limpian los archivos de origen para despojarlos de los elementos paratextuales que interfieren con su tratamiento. Esto puede aprovecharse para, sin trabajo adicional, preparar los textos para facilitar la automatización en la siguiente etapa. Esta consistente en tabular la información. Gracias a que las entidades constitutivas del texto han sido previamente marcadas, su distribución en las columnas correspondientes es trivial. Los versos así dispuestos se escanden considerando la fonología de las líneas dramáticas. Se emplea una biblioteca diseñada para este propósito, que se aplica a los versos uno tras otro, ajustando en tiempo real los parámetros para analizar cada uno de ellos, de acuerdo a su contexto métrico y dramático. Finalmente, se agrupan los versos según su rima y metro.

Todo esto tiene sentido como eslabones de una cadena. Sin embargo, al contrario que con

una real, los eslabones constitutivos de este proceso tienen utilidad por sí mismos. Las posibilidades dentro del ámbito de la lingüística que ofrece la división silábica de acuerdo a criterios fonológicos son formidables. Lo mismo puede decirse de la transcripción fonológica que presenta los resultados tanto en Unicode como en SAMPA, que abre un abanico de posibilidades en otros campos relacionados con el estudio de la lengua. Respecto a la escansión métrica, de la misma manera que se usa integrada en un programa que evalúa el contexto en tiempo real y adapta los parámetros de entrada en consecuencia, también puede aplicarse a tareas simples de metro fijo, como analizar poemas isométricos o versos sueltos, sin más expediente que crear el objeto correspondiente con los valores por defecto de la clase.

No menos interesante es la polivalencia de la propia secuencia de trabajo. La recopilación y anotación de los corpus incluye una inspección visual de los resultados. Durante esta, saltan a la vista a menudo metros extraños dentro de una determinada tirada o cambios estróficos breves. Sin embargo, con frecuencia, esto no evidencia fallos en la en la escansión de versos y su agrupación, pues el programa devuelve el resultado correcto de acuerdo a los datos de entrada. Por el contrario, la causa puede hallarse en las propias fuentes. En nuestro caso, dado que solo se consideraron textos fijados, de acuerdo a criterios filológicos estrictos, esto delataba errores de composición, transmisión o edición que habían logrado zafarse del escrutinio del editor crítico. En otras palabras, encontraría uso como herramienta ecdótica auxiliar para poner de relieve inconsistencias a cambio de unos pocos minutos.

En definitiva, hemos aplicado las premisas de que los instrumentos sencillos encuentran utilidad en las más dispares situaciones y un trabajo complejo pueden concebirse como la combinación de otros simples. De esta manera, en lugar de atacar el problema como un bloque monolítico, lo hemos dividido en partes separadas, que hemos abordado mediante mecanismos sencillos, pero efectivos dentro de su rango de aplicación. El resultado es una caja de herramientas cuyas piezas, combinadas de la manera adecuada, permiten al proyecto cumplir su cometido. No obstante, al estar estos instrumentos ideados para llevar a cabo tareas mínimas y, por lo tanto, comunes a muchos otros ámbitos, pueden asimismo emplearse en otros entornos y recombinarse de otras maneras para alcanzar propósitos alternativos.

REFERENCIAS BIBLIOGRÁFICAS

Ehrlicher, H., & Lehmann, J. (2019). Datenerhebung als epistemologisches Labor: Überlegungen am Beispiel der virtuellen Forschungsumgebung *Revistas culturales 2.0*. En M. Huber, S. Krämer y C. Pias (Eds.), *Forschungsinfrastrukturen in den digitalin Geisteswissenschaften: wie verändern digitale Infrastrukturen die Praxis der Geisteswissenschaften? „Digitalität in den Geisteswissenschaften“)* (pp. 40-57). CompaRe.

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., & Trilcke, P. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. En *Proceedings of DH2019: "Complexities", Utrecht, July 9-12*. Utrecht University. https://doi.org/10.5281/zenodo.4284002

Free Software Foundation (2022). Licenses. *GNU Operating System*. https://www.gnu.org/licenses/licenses.html

García Adánez, I. (2005). Reflexiones sobre la composición fonética y el lenguaje como material de la composición. En A. M. Reboul, A. Gimber y A. I. Fernández Valbuena (Eds.), *Palabra y música* (pp. 95-104), Universidad Complutense de Madrid.

Gervás, P. (2000). A Logic Programming Application for the Analysis of Spanish Verse. *Computational Logic - CL 2000, First International Conference*, 1330-1344. https://doi.org/10.1007/3-540-44957-4_89

Ilsemann, H. (1998). *Shakespeare Disassembled: Eine quantitative Analyse der Dramen Shakespeares*. Peter Lang.

Kroll, S., & Sanz-Lázaro, F. (2022). Romances teatrales entre Mira de Amescua, Calderón y Lope: ritmo, asonancia y cuestiones de autoría. *Revista de Humanidades Digitales*, *7*, 1-18. https://doi.org/10.5944/rhd.vol.7.2022.31620

Kroll, S., & Sanz-Lázaro, F. (2023). Ritmo, autoría y género: nuevas perspectivas sobre teatro lopesco desde las humanidades digitales. *Anuario Lope de Vega*, 29, 351-375. https://doi.org/10.5565/rev/anuariolopedevega.491

Marco, G., Rosa, J. de la, Gonzalo, J., Ros, S., & González-Blanco, E. (2021). Automated Metric Analysis of Spanish Poetry: Two Complementary Approaches. *IEEE Access, 9, 51734-51746*. https://doi.org/10.1109/ACCESS.2021.3069635

Marco Remón, G., & Gonzalo Arroyo, J. (2021). Escansión automática de poesía española sin silabación. *Procesamiento del Lenguaje Natural*, *66*, 77-87. https://doi.org/10.26342/2021-66-6

Moretti, F. (2000). Conjectures on World Literature. *New Left Review*, *1*, 54-68. https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature

Navarro-Colorado, B. (2015). A Computational Linguistic Approach to Spanish Golden Age Sonnets: Metrical and Semantic Aspects. *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, 105-113. https://doi.org/10.3115/v1/W15-0712

Navarro-Colorado, B. (2017). A Metrical Scansion System for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities*, *33*(*1*), 112-127. https://doi.org/10.1093/llc/fqx009

Navarro-Colorado, B., Ribes Lafoz, M., & Sánchez, N. (2016). Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. En N. C. (Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk y S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Presotto, M. (2023). Ecdótica y Humanidades Digitales en el estudio del Teatro clásico español: una experiencia. *Revista de Humanidades Digitales*, *8*, 1-13. https://doi.org/10.5944/rhd.vol.8.2023.37267

Python Software Foundation. (2023). *PyPI · The Python Package Index*. https://pypi.org/

Romanska, M. (2015). Drametrics: What Dramaturgs Should Learn from Mathematicians. En M. Ro-

manska (Ed.), *The Routledge companion to dramaturgy* (pp. 438-447). Routledge.

Rosa, J. de la, & Pérez, Á. (2020). Linhd-postdata/edfu: First Version of EDFU Syllabification Corpus. https://zenodo.org/record/3898684

Rosa, J. de la, Pérez, Á., Hernández, L., Ros, S., & González-Blanco García, E. (2020). Rantanplan, Fast and Accurate Syllabification and Scansion of Spanish Poetry. *Procesamiento del Lenguaje Natural*, *65*, 83-90. https://doi.org/https://doi.org/10.26342/2020-65-10

Rosa, J. de la, Pérez Pozo, Á., Ros, S., &González-Blanco García, E. (En prensa). ALBERTI, a Multilingual Domain Specific Language Model for poetry Analysis. En *SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing*.

Real Academia Española y Asociación de Academias de la Lengua Española [RAE y ASALE]. (2010a). *Nueva gramática de la lengua española. Manual*. Espasa.

Real Academia Española y Asociación de Academias de la Lengua Española [RAE y ASALE]. (2010b). *Ortografía de la lengua española*. Espasa.

Sánchez Jiménez, A. Acentos contiguos en los romances de la *Arcadia* (1598), de Lope de Vega. *Atalanta. Revista de las letras barrocas*, *5*(1), 5-61. http://doi.org/10.14643/51A

Sanz-Lázaro, F. (2022a). *Silabeador*. https://github.com/fsanzl/silabeador

Sanz-Lázaro, F. (2022b). *Fonemas*. https://github.com/fsanzl/fonemas

Sanz-Lázaro, F. (2023a). libEscansión. https://github.com/fsanzl/libEscansion

Sanz-Lázaro, F. (2023b). Planteamientos digitales del drama aurisecular: automatización del análisis métrico de obras teatrales. En I. González Cabeza, É. Redruello Vidal y R. de la Varga Llamazares (Eds.), *La escritura en el tiempo: investigaciones en torno a la literatura hispánica* (pp. 109-126). Universidad de León.

Sanz-Lázaro, F. (2023c). *txt2tei*. https://github.com/fsanzl/txt2tei