Edición digital enriquecida: un modelo de anotación multinivel para poesía del Siglo de Oro

Borja Navarro Colorado
Universidad de Alicante
borja@dlsi.ua.es

Resumen
En este capítulo se presenta un modelo general para la anotación multinivel de corpora de texto literario. Por multinivel se hace referencia a la combinación, en un mismo corpus, de información de diferentes niveles de descripción lingüística o literaria, desde datos relacionados con palabras o sílabas, hasta cuestiones temáticas, textuales o pragmáticas. El objetivo final de un corpus de estas características es fijar un posible análisis literario, por lo que se considera como una edición digital enriquecida. Se defienden cuatro características que un corpus de texto literario debe cumplir: interoperabilidad, perspectivismo, unidad y claridad/sencillez. Se da cuenta de los principales problemas de formalización en un corpus multinivel de este tipo: la combinación de diferentes formalismos de representación y, en el caso de XML, el problema de un anidamiento incorrecto. Finalmente se propone un modelo para un corpus de poesía del Siglo de Oro.

Palabras clave
corpus multinivel; poesía; métrica; sintaxis; anotación; edición digital; XML; TEI.

Abstract
Enriched Digital Edition: a Multilevel Annotation Model for Golden-Age Spanish Poetry.

This paper presents a general model for the multilevel annotation of a literary corpus. Multilevel refers to the combination of information from different linguistic or literary levels in the same corpus: from word related data to thematic, textual or pragmatic questions. The objective is to fix a possible literary analysis. To be considered an enriched digital edition, an annotated corpus must meet four

characteristics: interoperability, perspectivism, unity and clarity/simplicity. The main formalization problems are discussed: the combination of different representation formalisms and, in the case of XML, the improper nesting. Finally, a model for a corpus of poetry from the Spanish Golden-Age is proposed.

Keywords
multilevel corpus; poetry; meter; syntax; annotation; digital edition; XML; TEI.

Procesamiento del lenguaje natural y los estudios literarios computacionales[1]

Una de las áreas de los llamados estudios literarios computacionales (CLS por sus siglas en inglés)[2] es la aplicación de herramientas de procesamiento del lenguaje natural (PLN) e inteligencia artificial (IA) al estudio del texto literario. Esta aplicación presenta diversos problemas.

En primer lugar, las herramientas de PLN, salvo casos excepcionales, se crean para procesar un texto moderno estándar. El análisis automático del texto literario es mucho más complejo por dos motivos principalmente. Por un lado porque el texto literario suele ser texto antiguo, con variantes históricas del idioma (aunque esté modernizado) para las que las herramientas de PLN no están preparadas. A esto se le une la desautomatización propia del estilo literario, por la cual se exprime al máximo la capacidad expresiva del idioma y se buscan formas lingüísticas novedosas.

Por otro lado, las herramientas de PLN no suelen estar diseñadas para extraer información literaria. Éstas analizan aquellos aspectos lingüísticos que la comuni-

1. Trabajo parcialmente financiado por el Ministerio de Ciencia e Innovación a través del proyecto "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00): MCIN/AEI/10.13039/501100011033/ y "FEDER Una manera de hacer Europa"; y por la Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) a través del Proyecto: NL4DISMIS: Tecnologías del Lenguaje Natural para lidiar con la desinformación (CIPROM/2021/021).
2. Para una visión general de los métodos actuales en los estudios literarios computacionales, véase Schöch *et at.* eds. (2023).

dad de PLN considera relevantes, como categorías gramaticales, sintaxis, sentimientos o entidades nombradas, entre otras.[3] Esta información solo resulta en parte de interés para los estudios literarios. Son pocas las herramientas de PLN capaces de extraer rasgos específicos del texto literario, como puede ser la métrica (Navarro Colorado 2018a, de la Rosa *et al.* 2020, Haider 2021), figuras retóricas y metáforas novedosas (Kesarwani *et al.* 2017), análisis de personajes, detección de eventos narrativos y otros aspectos narratológicos (Piper *et al.* 2021), etc.

Para poder aplicar, por tanto, análisis computacional al texto literario es necesario desarrollar herramientas específicas capaces, por un lado, de interpretar el texto literario en su complejidad histórica y expresiva, y por otro, que sean capaces de extraer información de interés para los estudios literarios.

Para desarrollar estas herramientas hay que aplicar las técnicas propias del PLN. El paradigma técnico dominante para el desarrollo de herramientas de PLN en los últimos 30 años ha sido el aprendizaje automático o *machine learning* (Jurafsky y Martin 2023),[4] y en concreto el aprendizaje automático supervisado.[5] A diferencia de los métodos tradicionales en los que se indica a la máquina cómo analizar un texto mediante reglas de programación, en el paradigma de aprendizaje automático supervisado es la propia máquina la que infiere cómo analizar textos a partir de muestras textuales anotadas por humanos. Estas muestras textuales son los corpora anotados, que como se puede comprobar, tienen un papel básico en PLN, bien sea como corpus de aprendizaje en los métodos supervisados, bien como corpus de evaluación en cualquier caso (Pustejovsky y Stubbs 2012, Ide y Pustejovsky (eds.) 2017).

Con esta situación, para los estudios literarios computacionales en general, y para el desarrollo de herramientas de PLN específicas para texto literario en particular, es necesario desarrollar corpora de textos literarios anotados con información tanto lingüística como propiamente literaria.

La anotación de corpus se puede considerar además un modelo de análisis lingüístico y literario en sí mismo. Anotar un corpus es una tarea de interpretación

3. Véase Jurafsky y Martin (2023), en especial los capítulos 17 a 28, centrados en el análisis computacional de estructuras lingüísticas.

4. En el PLN actual sin embargo, dominan las técnicas neuronales y el desarrollo de grandes modelo de lenguaje (large language models o LLM), caracterizados precisamente por no necesitar corpora anotados a mano. Sin embargo, tanto para adaptar esos grandes modelos de lenguaje a tareas concretas (como las que se requieren en CLS), el llamado fine-tunning, como para las técnicas de prompting basadas en ejemplos (few-shots prompting), los corpora anotados sí son necesarios.

5. Las técnicas de aprendizaje automático se agrupan en dos grandes familias: las técnicas supervisadas y las no supervisadas. Las supervisadas parten de un corpus anotado, mientras que las no supervisadas parten de un corpus sin ningún tipo de marca ni anotación. En este trabajo me centro en la aproximación supervisada, que es la que más relevancia tiene en los corpora anotados. En los estudios literarios también se han aplicado técnicas no supervisadas, como por ejemplo *Topic Modeling* (Navarro Colorado 2018*b*). Sobre la aplicación de aprendizaje automático a los estudios literarios computacionales, véase Hatzel *et al.* (2023).

y estudio de la lengua literaria a partir del análisis de muestras textuales. Entre otros aspectos requiere definir un modelo teórico general, especificar qué información se va a marcar, formalizar esa información mediante un lenguaje de marcas, determinar el proceso de anotación, detectar casos complejos (normalmente problemas de ambigüedad) y determinar heurísticas para su correcta interpretación y anotación, entre otros aspectos (Ide y Pustejovsky (eds.) 2017). Aplicado a un texto literario, es una tarea hermenéutica propia de los estudios literarios (Meister 2022).

La anotación de corpus como paradigma de análisis lingüístico-literario no está exenta de problemas. Sin entrar en detalles, comentaré dos aspectos. Por un lado, la anotación tiene sus límites porque no se puede asumir que se pueda anotar todo tipo de información lingüístico-literaria. Por otro, toda formalización supone siempre un proceso de simplificación, y por tanto supone cierta pérdida en la precisión de análisis; en especial los casos más complejos (aquellos que se desvían de la norma general), que se obvian o se simplifican demasiado para ajustarlos al modelo general (McShane y Nirenburg 2001: 52-55). Como principal ventaja, la anotación del corpus permite automatizar el análisis literario y, así, realizar análisis a gran escala de amplios corpora.

Desde la óptica de los estudios literarios, un corpus anotado es lo que vamos a denominar una "edición digital enriquecida", en tanto que edición digital de una o más obras en las que no solo se ha fijado el texto, sino que también se ha establecido una posible interpretación de las obras a uno o varios niveles de análisis: métrica, sintáctica, semántica, temática, etc. Y se ha realizado además con ánimo de ser análisis de referencia para futuras investigaciones del fenómeno literario tratado.

En este trabajo presentamos un modelo general para la anotación multinivel de corpora de texto literario. Por multinivel me refiero a un modelo de anotación que permita combinar información de diferentes niveles de descripción lingüística o literaria, de ahí el nombre de "multinivel": desde datos relacionados con palabras o sílabas, hasta cuestiones temáticas, textuales o pragmáticas. En la siguiente sección se expondrán los requisitos que considero debe tener un corpus de texto literario anotado para considerarse edición digital enriquecida; luego comentaré los principales problemas de formalización que presenta el desarrollo de un corpus multinivel, y finalmente se propondrá un modelo general para un corpus de poesía del Siglo de Oro.

Requisitos generales para una edición digital enriquecida

Desarrollar un corpus anotado es una tarea compleja que requiere tiempo y esfuerzo. Se debe abordar siempre con la intención de conseguir una anotación útil y válida durante mucho tiempo. Así como en una edición crítica se busca fijar un texto de referencia para su lectura y análisis posterior por parte de la comunidad científica, una edición enriquecida busca marcar en el texto un aná-

lisis de referencia que sea útil para los estudios literarios durante los siguientes años. No me refiero con esto a establecer una única interpretación del texto literario (más sobre este punto luego), sino a establecer un análisis que sirva de referencia para otros análisis, bien como soporte para investigaciones relacionadas, bien para proponer análisis alternativos, o bien como corpus de aprendizaje para sistemas automáticos.

Para conseguir, por tanto, que un corpus anotado como el aquí planteado sea útil, fiable y válido para los estudios literarios, de manera que pueda ser considerado una edición digital enriquecida, debe cumplir cuatro requisitos básicos:

- interoperabilidad,
- perspectivismo,
- unidad
- sencillez y claridad.

A continuación comentaré cada uno de ellos.

Interoperabilidad

El concepto de interoperabilidad proviene de los sistemas de información y su capacidad para utilizarse de forma conjunta, compartir datos e intercambiar información.[6] Así, entendiendo un corpus anotado como un sistema de información, debe estar diseñado de tal manera que la información contenida (la anotación) pueda ser compartida y reutilizada por otros corpora y/o por sistemas automáticos y herramientas de PLN.

Este rasgo recae sobre todo en la formalización. El formato de anotación debe ser lo más compatible posible, incluyendo aquí tanto el lenguaje de marcado utilizado, como la sistematización de la información realizada y la representación de ésta mediante etiquetas. Esto asegura la reutilización de la información anotada, su complementariedad con otros corpora y, en general, el intercambio de información con diferentes sistemas automáticos.

Para ello es necesario utilizar lenguajes y modelos de marcado estándar. En humanidades digitales y estudios literarios computacionales, el lenguaje de marcas más utilizado es el eXtensible Markup Language o XML usado según las recomendaciones de la *Text Encoding Initiative* o TEI (TEI Consortium 2023). Sin embargo, no siempre XML-TEI es la mejor opción. En la siguiente sección se expondrán sus límites y algunos problemas que pueden surgir al anotar un corpus multinivel con este formato estándar.

6. Véase cómo define interoperabilidad el *Diccionario panhispánico del español jurídico*: https://dpej.rae.es/lema/interoperabilidad (19 de septiembre de 2023).

Perspectivismo

En el área del PLN, la calidad de un corpus de aprendizaje se determina según la consistencia de la anotación (Pustejovsky y Stubbs 2012, Ide y Pustejovsky 2017): en qué medida, un mismo fenómeno lingüístico es anotado de la misma manera por dos personas diferentes. Se busca el máximo consenso en la anotación del corpus para que la máquina, al final, aprenda aquello en lo que los humanos están de acuerdo. El desacuerdo entre los anotadores se penaliza, pues se considera que hay un error en la anotación (bien sea en la formalización, en la sistematización, en las etiquetas, o en cualquier otro aspecto).

Este planteamiento es apropiado en la anotación lingüística de corpora en la medida en que, con un mismo modelo teórico, es viable llegar a un acuerdo en la interpretación de cuestiones morfológicas o sintácticas. Sin embargo, en las últimas décadas el PLN ha ido abordando cuestiones semánticas y más subjetivas donde el acuerdo interpretativo entre los anotadores cada vez es más complejo, incluso partiendo de los mismos modelos teóricos. En los últimos años, al tratar con temas tan subjetivos como la expresión de opiniones, emociones y sentimientos, o la detección de la ironía, se ha replanteado este modelo de calidad de un corpus anotado basado en el acuerdo entre anotadores.

En este contexto ha surgido un nuevo paradigma en anotación de corpora denominado "perspectivismo".[7] Este paradigma considera que el desacuerdo entre anotadores no es tanto un error como una fuente valiosa de información sobre el fenómeno anotado. El desacuerdo entre anotadores manifiesta formas diferentes de interpretar un mismo fenómeno lingüístico o (en nuestro caso) literario. Este desacuerdo es relevante porque, en primer lugar, está mostrando problemas en la anotación y por tanto temas relevantes para ser estudiados e investigados (¿a qué se debe el desacuerdo?, ¿es un fallo en el modelo teórico general, en las etiquetas, en el proceso?, ¿es un caso nuevo que obliga a replantear el modelo, la definición del fenómeno?, etc.); y en segundo lugar es relevante porque si entre las personas expertas no hay acuerdo, las herramientas computacionales derivadas del corpus deben aprender también que ese fenómeno se puede interpretar (es decir, anotar) de diversas maneras, que no hay una interpretación única (Cabitza *et al.* 2023).

El perspectivismo, por tanto, hace referencia a la posibilidad de anotar los mismos fenómenos y los mismos textos de diferentes maneras, desde supuestos, modelos o perspectivas diferentes. Toda anotación implica un proceso de interpretación, y se debe asumir que siempre puede haber diversas interpretaciones. Un corpus anotado debe ser consistente consigo mismo, es decir: una

7. Véase *The Perspectivist Data Manifesto* en https://pdai.info/ (19 de septiembre de 2023)

misma persona debe anotar los mismos fenómenos de la misma manera a lo largo de todo el corpus. Pero no tiene por qué ser consistente con la anotación de otra persona.

Desde este punto de vista, una edición crítica enriquecida, en tanto que texto literario anotado, debe ser considerada siempre como una propuesta de análisis. La anotación fija una interpretación del texto que se presenta como referencia para otros estudios, pero no como una interpretación única y definitiva: se asume que puede haber (y debe haber) anotaciones y análisis alternativos que enriquezcan la interpretación. Si en otras áreas del PLN el perspectivismo es una opción, para la anotación del texto literario, dado su carácter plurisignificativo, es una necesidad. Piénsese, por poner un caso extremo (pero no extraño), las diferentes relaciones sintácticas y, con ello, las diferentes interpretaciones que se pueden establecer en estos famosos versos de Quevedo:

> Cerrar podrá mis ojos la postrera
> sombra que me llevare el blanco día,
> y podrá desatar esta alma mía
> hora a su afán ansioso lisonjera:
> (Quevedo, *Amor constante más allá de la muerte*, vv. 1-4)[8]

Unidad

La unidad hace referencia a la unidad textual del corpus. Un texto literario anotado con información diversa a diferentes niveles de descripción lingüística puede llegar a ser un entramado complejo de ficheros e información. Pero no hay que perder nunca de vista la unidad del corpus: toda la anotación responde a un único texto y todos los elementos anotados deben estar relacionados.

Un poema, por ejemplo, anotado con información métrica y con información categorial debe estar creado de tal manera que, aunque sean aspectos relacionados con unidades diferentes (la sílaba para la información métrica y la palabra para la información categorial), quede claro que son rasgos de un mismo texto y que están relacionados.

Dada la diversidad de información y la heterogeneidad de unidades lingüísticas, establecer estas relaciones no es tarea fácil. En la siguiente sección se expondrán problemas derivados de este aspecto y vías de solución.

8. Jauralde Pou (1997) repasa los diferentes problemas interpretativos del poema y las lecturas clásicas. En concreto, para los versos 3-4, que considera "extremadamente difíciles" (pág. 94), destaca "la extraña utilización de *hora*; la concordancia de *lisonjera* y la referencia de *su*." (pág. 95). Todos estos son problemas sintácticos. "Hora", por ejemplo, suele considerarse sujeto de "podrá desatar" (Torre, 2004), pero no es una lectura evidente.

Sencillez y claridad

Precisamente por la complejidad que tiene un corpus anotado, se debe procurar que el modelo formal (el código, las marcas y las etiquetas utilizadas en la anotación) sea lo más claro posible para la persona que realiza la anotación manual. Con independencia del uso de editores de anotación como CATMA,[9] INCEpTION[10] u otros, un código claro permite a las personas modificarlo directamente. El objetivo es conseguir, en la medida de lo posible, un marcado claro, sencillo y explícito en la línea de las recomendaciones de código claro para lenguajes de programación como Python.[11]

No quiero expresar con esto un rechazo absoluto al uso de editores de anotación. Más bien quiero expresar que el uso de editores de anotación debe ser siempre una opción del anotador, pero nunca una obligación. Los editores mediatizan la anotación y limitan la capacidad expresiva de los lenguajes de marcado.[12] Por ello siempre hay que mantener la opción de anotar directamente en código (con editores de código para asegurar la consistencia sintáctica y semántica), y ello implica que el marcado, el código, debe ser inteligible para un humano. Dado que en toda anotación se hará antes o después un trabajo manual, considero que hay que dar preferencia antes a la inteligibilidad del marcado que a su optimización computacional.

En resumen, de una edición digital enriquecida se espera que fije una interpretación de referencia del texto útil para los estudios computacionales, pero asumiendo siempre interpretaciones y anotaciones alternativas; que sea compatible con otros corpora y sistemas automáticos; que todos los niveles de anotación estén relacionados para mantener la unidad textual del corpus, y que el código sea, en la medida de lo posible, claro e inteligible para una persona experta.

Problemas formales en la anotación de un corpus multinivel.

Conseguir estas características implica resolver diversos problemas. En esta sección comentaré cuestiones derivadas de la formalización, la unidad y la claridad del código. A partir de ello, en la siguiente sección propondré un modelo general para poesía del Siglo de Oro.

9. https://catma.de/ (19 de septiembre de 2023)
10. https://inception-project.github.io/ (19 de septiembre de 2023)
11. Véase, por ejemplo, "Zen of Python" en *The Hitchhiker's Guide to Python*: https://docs.python-guide.org/writing/style/#zen-of-python
12. El editor CATMA, por ejemplo, siguiendo las propias recomendaciones de TEI, separa el texto de la anotación en dos ficheros diferentes. El fichero con las etiquetas tiene referencias a las posiciones del fragmento de texto al que se refieren la etiquetas. Si bien este modelo es óptimo desde un punto de vista computacional, la anotación resultante resulta opaca para un humano.

Se comentó en la sección anterior que para conseguir la máxima interoperabilidad es necesario marcar el corpus con lenguajes de marcado de uso común y, si existe, seguir un estándar de anotación. En humanidades digitales, la opción más estandarizada es sin duda alguna el formalismo XML-TEI. Ahora bien, al afrontar la creación de un corpus multinivel, hay dos cuestiones generales a las que hay que dar respuesta: en primer lugar, cómo anotar aquellos fenómenos que no se pueden representar con XML-TEI o para los que hay un estándar alternativo; en segundo lugar, cómo resolver el problema de anidamiento que se puede producir entre los diferentes niveles de anotación.

Límites de XML-TEI

Si bien XML es un lenguaje de marcas muy flexible, que se puede adaptar a prácticamente cualquier tipo de información que se quiera representar, no siempre es la mejor opción. En anotación lingüística de corpora hay un caso en el que XML no es el lenguaje de marcas más utilizado: la sintaxis.

El modelo teórico más implantado hoy en sintaxis computacional es el modelo de dependencias. Mediante este modelo se marcan las relaciones de dependencia entre las palabras (lo que en la gramática tradicional se denominan funciones sintácticas: sujeto, objeto directo, etc.). La relación es siempre entre dos palabras, una es la palabra núcleo y otra la palabra dependiente. Todas las relaciones de dependencia entre las palabras de una oración forman al final el árbol sintáctico. Véase Fig. 1:

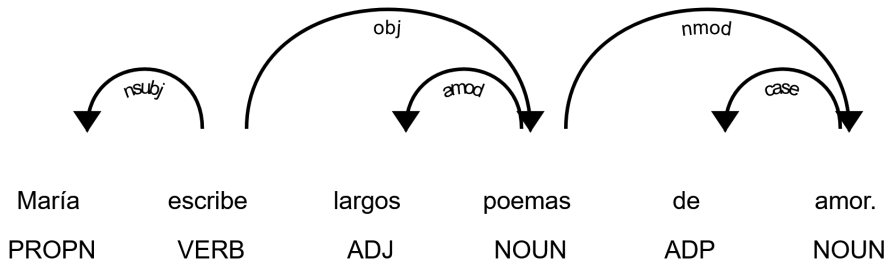

Fig. 1: Ejemplos de análisis sintáctico de dependencias.

Desde un punto de vista formal, un árbol de dependencias se representa como un grafo dirigido formado por un conjunto de vértices y un conjunto de arcos. Los vértices son las palabras de la oración y los arcos dirigidos son las relaciones de dependencia, que van siempre de un vértice (la palabra que actúa de núcleo) a otro vértice (la palabra dependiente de ese núcleo) (Jurafsky y Martin 2023).

El problema para la anotación de corpora con información sintáctica es que, si bien es posible representar grafos en XML, éste no suele ser el formalismo estándar. Hay otros modos más comunes de representar formalmente un grafo. En el caso concreto del análisis sintáctico, dentro del mundo del PLN el formato más común es el formato CONLL (Buchholz y Marsi 2006).[13] La Tabla 1 muestra un ejemplo del análisis sintáctico del verso de Lope de Vega "Un soneto me manda hacer Violante" con un formato CONLL simplificado:

1	Un	Uno	DI0MS0	2	spec
2	soneto	soneto	NCMS000	5	obj
3	me	me	PP1CS00	4	obl
4	manda	mandar	VMIP3S0	0	sentence
5	hacer	hacer	VMN0000	4	obj
6	Violante	violante	NP00000	4	suj
7	.	.	Fp	4	f

Tabla 1: Análisis sintáctico de dependencias en formato CONLL (simplificado).

Como se puede ver, el formato CONLL es un formato tabular, donde las líneas representan cada palabra y las columnas su información morfosintáctica. La primera columna es el número identificativo de la palabra; y las siguientes tres columnas la información morfológica: el *token* tal y como aparece en el texto, su lema y la etiqueta categorial y morfológica. Finalmente, en las columnas 5 y 6 aparece la información sintáctica. La columna 5 es el número de la palabra de quien depende (su núcleo) y la 6 el tipo de dependencia.[14]

Así, la raíz de este árbol sintáctico es la palabra 4 ("mandar"), de la cual depende la palabra 6 ("Violante") con una dependencia tipo "sujeto" (suj) y la palabra 5 ("hacer") con una dependencia tipo "objeto" (obj), etc.

El siguiente código muestra este mismo grafo en anotación XML siguiendo las recomendaciones TEI:[15]

```
<graph type="directed" xml:id="RDG1">
<node n="1"><label>
      <w lemma="uno" pos="DI0MS0" msd="DI">Un</w></label>
</node>
<node n="2"><label>
```

13. Véase https://universaldependencies.org/format.html (19 de septiembre de 2023.)
14. En este caso, las etiquetas morfológicas y sintácticas están basadas en el modelo de Dependencias Universales. Véase https://universaldependencies.org/ (19 de septiembre de 2023).
15. Véase el capítulo 19 de la guía TEI "Graphs, Networks, and Trees": https://tei-c.org/release/doc/tei-p5-doc/en/html/GD.html

```
        <w lemma="soneto" pos="NCMS000" msd="NC">soneto</w></label>
    </node>
    <node n="3"><label>
        <w lemma="me" pos="PP1CS00" msd="PP">me</w></label>
    </node>
    <node n="4"><label>
        <w lemma="mandar" pos="VMIP3S0" msd="VMI">manda</w></label>
    </node>
    <node n="5"><label>
        <w lemma="hacer" pos="VMN0000" msd="VMN">hacer</w></label>
    </node>
    <node n="6"><label>
        <w lemma="Violante" pos="NP00000" msd="NP">Violante</w></label>
    </node>

    <arc from="#1" to="#2">
        <label>spec</label>
    </arc>
    <arc from="#2" to="#5">
        <label>obj</label>
    </arc>
    <arc from="#3" to="#4">
        <label>obl</label>
    </arc>
    <arc from="#4" to="#0">
        <label>sentence</label>
    </arc>
    <arc from="#5" to="#4">
        <label>obj</label>
    </arc>
    <arc from="#6" to="#4">
        <label>subj</label>
    </arc>
    </graph>
```

En este caso, primero se codifican los nodos del arco con la etiqueta <node> y luego los arcos dirigidos con la etiqueta <arc>, que especifica la relación entre dos nodos o palabras (@from y @to).

Como se puede observar, la codificación CONLL es más inteligible para un humano y es también óptima para su tratamiento computacional.

La cuestión, por tanto, a la hora de desarrollar una edición digital enriquecida multinivel, es decidir si para marcar la información sintáctica se utiliza XML, como presumiblemente se habrá utilizado en otros niveles, o utilizar otro formalismo diferente como CONLL, de uso común y considerado el estándar *de facto* en la comunidad científica.

Esto mismo podría ocurrir con otro tipo de información. Así, para la anotación multinivel de corpora lingüísticos hay propuestas en los dos sentidos. Por

ejemplo, PAULA-XML[16] (Chiarcos *et al.* 2008) o SALT[17] (Zipser y Romary 2010), al igual que TEI, proponen un único formalismo para marcar cualquier tipo de información en el corpus. El primer caso está basado en XML, mientras que el segundo está basado en grafos. De esta manera toda la información está unificada en un único formalismo. Sin embargo, otros corpora, como el corpus GUM[18] (Zeldes 2017), utilizan un formalismo diferente para cada tipo de información o nivel de anotación, según sea el estándar o el formato más común. Así, en el corpus GUM la estructura general de cada documento está marcada en XML-TEI, pero luego utilizan CONLL para marcar la información sintáctica.

Dentro de los estudios literarios computacionales, esta opción de mantener el formato más usado para cada tipo de información o nivel de representación es la que se utiliza en BookNLP:[19] un conjunto de herramientas de PLN entrenadas específicamente para procesar texto literario en inglés.[20] Por el contrario, algunos editores de marcado como CATMA[21] o TEI:TOK,[22] utilizan XML-TEI como formato nativo y cualquier tipo de información que se pueda anotar se representa con este formalismo único.

Con esta situación, no se puede considerar una opción mejor que otra. Si se opta por utilizar diferentes lenguajes de marcado, se corre el riesgo de perder la unidad del corpus y que éste quede como una acumulación de ficheros. Si se opta por utilizar XML para toda la anotación, puede resultar un código final oscuro; aparte del problema de anidamiento que se comentará después.

En cuanto a los límites de TEI, si bien cuenta con infinidad de etiquetas, no incluye toda la información que pueda interesar para los estudios literarios computacionales. Pero esto realmente no es un problema porque TEI ya está diseñado para que se pueda ampliar con la especificación de etiquetas propias, aunque no sea una práctica recomendada.[23]

Anidamiento

Antes se comentó que la información que se puede anotar en un corpus multinivel es muy variada, desde información lingüística como lemas y categorías gramatica-

16. https://github.com/korpling/paula-xml (19 de septiembre de 2023)
17. https://corpus-tools.org/salt/ (19 de septiembre de 2023)
18. https://gucorpling.org/gum/ (19 de septiembre de 2023)
19. https://github.com/booknlp/booknlp (19 de septiembre de 2023)
20. Junto al análisis lingüístico, realiza un análisis propio del texto literario como la detección automática de personajes y eventos narrativos, o la detección de estilo directo, entre otros aspectos.
21. https://catma.de/ (19 de septiembre de 2023)
22. http://teitok.corpuswiki.org/ (19 de septiembre de 2023)
23. Véase la sección "Customization" de la web de TEI: https://tei-c.org/guidelines/customization/. Para recursos en español, véase la web TTHUB https://tthub.io/ (19 de septiembre de 2023) y en especial Allés-Torrent *et al.* (2022).

les, frases hechas, sintaxis, referencias a entidades (*named entities*), emociones y opiniones, relaciones anafóricas, estructuras argumentales y roles semánticos, significado léxico, etc. (Jurafsky y Martin 2023); hasta información literaria como los personajes de novelas y las relaciones entre ellos, intervenciones y estilo de discurso, eventos narrativos, métrica y ritmo, sentimientos, temas recurrentes, similitudes estilométricas, etc. (Piper *et al.* 2021, Schöch et at. eds. 2023, entre otros).

Si se opta por utilizar XML como único formalismo para representar información diversa en un único corpus, surge el problema del cruce de ramas, es decir, un anidamiento erróneo de los elementos XML.

La representación que hace XML de la información es una representación arbórea y anidada. De una etiqueta madre (la principal es *root*) pueden depender una, dos o más etiquetas, de las cuales a su vez pueden depender otras etiquetas, etc. Así, un texto marcado en XML es un árbol descendente, donde cada etiqueta de un nivel tiene anidadas las etiquetas del siguiente nivel.[24]

Por ejemplo, para marcar la estructura básica de una novela en XML, TEI propone usar una etiqueta "text", de la que depende una etiqueta "body", de la que a su vez depende una o más etiquetas "chapter", de la cual dependen una o más etiquetas "p" (párrafo), etc. Se forma, así, una estructura arbórea que se puede recorrer de manera descendente o ascendente.

Esta estructura arbórea establece relaciones de anidamiento entre etiquetas. Esto implica que si se introduce una etiqueta de apertura <A> y dentro otra etiqueta de apertura <B>, no se puede cerrar la etiqueta <A> hasta que se haya cerrado la etiqueta anidada B. Si así se hiciera, el XML resultante estaría mal formado por un error de anidamiento y no se podría procesar.

Este error puede aparecer cuando se quiere anotar en un mismo fichero información a diferentes niveles de descripción, porque cada uno puede depender de unidades lingüísticas diferentes. Así ocurre, por ejemplo, si se quiere anotar en un poema información métrica e información categorial. La unidad básica de la métrica es la sílaba métrica, mientras que la unidad básica de las categorías gramaticales es la palabra (el lema). En principio, son unidades compatibles porque una palabra está formada por sílabas, por lo que se puede asumir que siempre que finalice una palabra, finalizará también una sílaba. El problema surge cuando aparece una sinalefa.

Efectivamente, con la sinalefa nos hallamos ante un problema de anidamiento: al unirse la última sílaba de una palabra con la primera sílaba de la siguiente palabra en una única sílaba métrica, no se puede representar ambas unidades con XML en un mismo fichero. Si se cierra la etiqueta de palabra, se debe cerrar la etiqueta de sílaba (que está anidada) y por tanto se pierde el límite de la sílaba métrica. Si no se cierra, no se puede cerrar la unidad palabra y quedarían dos palabras marcadas como si fueran solo una.

24. http://www.w3.org/TR/xml/ (19 de septiembre de 2023)

En el verso de Lope de Vega mostrado en la Tabla 1 hay sinalefa entre las palabras “manda” y “hacer”. El siguiente ejemplo muestra las dos opciones de anotación (sílabas o palabras), pero no se pueden representar las dos a la vez:

```
<seg type=“syllable”>man</seg>
<seg type=“syllable”>daha</seg>
<seg type=“syllable”>cer</seg>

<w>manda</w>
<w>hacer</seg>
```

El estándar TEI propone algunas soluciones para poder tratar este problema (DeRose 2004).[25] En los siguientes párrafos las resumo brevemente con los siguientes versos de Lope de Vega como ejemplo. En este caso se produce un error de anidamiento entre la oración y el verso por el encabalgamiento del primer verso. Se muestra primero la anotación XML mal formada:

```
<l><s>Así Fabio lloraba. </s><s>Albania entonces </l>
<l>mirole, y quiso hablar, cerró los ojos, </l>
<l>y respondiole lo demás la muerte. </s></l>
```

Como se ve, no se puede marcar el final del primer verso (<l>) sin antes marcar el final de la oración (<s>), pero no se puede marcar el final de la oración porque ésta finaliza en el tercer verso.

La primera solución es hacer la anotación en ficheros diferentes. Así, cada uno tendrá su XML y no habrá problema de anidamiento. Con esta opción, sin embargo, se pierde la unidad del corpus: habría que idear métodos para combinar la información de cada fichero de manera automática.

```
<l>Así Fabio lloraba. Albania entonces</l>
<l>mirole, y quiso hablar, cerró los ojos,</l>
<l>y respondiole lo demás la muerte.</l>

<s>Así Fabio lloraba.</s>
<s>Albania entonces mirole, y quiso hablar, cerró
los ojos, y respondiole lo demás la muerte.</s>
```

Una segunda opción es marcar las unidades no con etiquetas dobles (una de apertura y otra de cierre, tipo <A> </A>), sino con etiquetas de elemento vacío <A/>, que aparecen solo una vez y marcan límites. La idea sería utilizar este tipo de etiqueta para marcar el inicio o el fin de la unidad que entra en conflicto con otra.

25. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html (19 de septiembre de 2023)

En este ejemplo, en vez de poner una etiqueta <s n=1> para marcar el inicio de la oración y otra </s> para marcar su final, se podría marcar solo el inicio de cada oración en un texto con <s n=1/>. El final de la oración se puede inferir por la etiqueta de inicio de la siguiente oración, o por la marca de fin de texto. Al ser solo una etiqueta, no se produce conflicto en el anidamiento de etiquetas. Este método no es tan explícito como el uso de dos etiquetas de inicio y cierre, y en muchas ocasiones no se puede utilizar, pero es una solución elegante y útil para algunos casos.

```
<l><s n="1"/>Así Fabio lloraba. <s n="2"/> Albania entonces </l>
<l>mirole, y quiso hablar, cerró los ojos, </l>
<l>y respondiole lo demás la muerte. </l>
```

Para la tercera opción es necesario que cada unidad tenga un identificador único (por ejemplo, si son oraciones, vale un simple @n, como n=1 para la primera oración, n=2 para la segunda, etc.). En este ejemplo, si una oración queda interrumpida por otra unidad, se cierra y se vuelve a abrir pero con el mismo identificador. De esta manera, gracias al identificador, queda marcado que la oración continúa:

```
<l><s n="1">Así Fabio lloraba.</s> <s n="2">Albania entonces </s></l>
<l><s n="2">mirole, y quiso hablar, cerró los ojos, </s></l>
<l><s n="2">y respondiole lo demás la muerte. </s></l>
```

El problema de esta opción es que se duplican las etiquetas y queda una anotación algo redundante. Si el corpus es muy amplio, se pueden generar al final ficheros muy grandes.

La última opción es el llamado *stand-off Markup*. La idea es tener en ficheros separados el texto por un lado (sin ningún tipo de marca, el texto puro) y en otros la anotación, con tanto ficheros como tipos de anotación se quieran incluir. Para marcar a qué fragmento de texto corresponde cada etiqueta, se utilizan referencias entre ficheros como punteros, enlaces, etc. Esta es la opción que utilizan editores como CATMA. Computacionalmente la opción es óptima pero, como se ha comentado antes, el resultado es una anotación opaca para las personas pues, en vez del texto, aparecen números de referencia.

El siguiente ejemplo es el formato de CATMA. En cada etiqueta el texto está referenciado con dos números (punteros): el primero indica la posición en el fichero de texto del primer carácter del fragmento de texto y el segundo la posición del último carácter.

```
<content>Así Fabio lloraba. Albania entonces
mirole, y quiso hablar, cerró los ojos,
y respondiole lo demás la muerte.</content>
```

```
<l><xi:include href="source.xml"
pointer="string-range(element(/1),0,35)"/></l>
<l><xi:include href="source.xml"
xpointer="string-range(element(/1),36,75)"/></l>
```

Como se puede observar, esta segunda opción es totalmente ininteligible para una persona. Eso sí, es muy eficiente desde un punto de vista computacional.

Qué solución adoptar depende de cada proyecto de anotación. En general, la opción 4 es la más eficiente desde un punto de vista computacional y la opción 2 la más elegante e inteligible, si bien no siempre es posible aplicarla. A partir de todos estos problemas, en la siguiente sección se mostrará un modelo general para la anotación multinivel de un corpus de poesía del Siglo de Oro.

Estudio de caso: un corpus multinivel de poesía del Siglo de Oro

En esta sección se aborda la anotación multinivel de un corpus de poesía del Siglo de Oro con el objetivo final de disponer de una anotación rica a modo de análisis de referencia que permita tanto profundizar en los estudios literarios como en el desarrollo de herramientas específicas de PLN para poesía española.

Los poemas están extraídos del *Corpus de sonetos del Siglo de Oro* desarrollado en el Proyecto ADSO (Navarro Colorado *et al.* 2016).[26] La edición de los textos es la modernización que realizó la Biblioteca Virtual Miguel de Cervantes,[27] más algunas correcciones y enmiendas. Además del texto, el corpus ya dispone de la anotación métrica realizada en el proyecto ADSO.

El modelo de anotación que aquí presentamos está pensado para poder marcar en principio cualquier tipo de información lingüística o literaria. Por ahora nuestro interés se centra en el estudio del hipérbaton como rasgo estilístico, para lo que se necesita marcar el corpus a tres niveles de descripción: métrico (el patrón métrico y rítmico de cada verso), léxico-categorial (la categoría gramatical de cada palabra) y sintáctico (las relaciones sintácticas entre las palabras, con independencia de su posición en el verso). En el futuro se espera poder anotar el corpus con otro tipo de información como entidades o lenguaje figurado.

El proceso de anotación, como es habitual, se desarrolla de manera semiautomática. Primero se realiza una anotación automática con herramientas de PLN, que luego es revisada, corregida y fijada de manera manual. En esta primera fase del proyecto estamos más interesados en fijar un análisis correcto que

26. https://gplsi.dlsi.ua.es/proyectos/adso/ (17 de enero de 2024).
27. https://www.cervantesvirtual.com/portales/biblioteca_del_soneto/ (19 de septiembre de 2023).

en disponer de gran cantidad de texto anotado, por lo que por ahora se está trabajando con pocos sonetos.

Buscando la máxima interoperabilidad, el lenguaje de marcado utilizado es el XML siguiendo, en la medida de lo posible, las recomendaciones del estándar TEI. Esta elección no representa un problema para la anotación métrica y categorial, que está bien definida en TEI y para la que XML es un lenguaje de marcado apropiado.

Sin embargo, como se ha comentado en secciones anteriores, la representación de la información sintáctica en XML-TEI no es una representación estándar ni común. Para conseguir la máxima interoperabilidad es mejor utilizar el formalismo CONLL, al igual que hacen otros corpora multinivel como el corpus GUM (Zeldes 2017): este es el formato que utiliza la herramienta de análisis sintáctico utilizado y es más inteligible para la anotación manual. Para mantener la unidad del corpus, se utilizan identificadores, como se comentará luego.

Con la anotación métrica y léxico-categorial, ambas en XML, se puede producir un problema de anidamiento comentado anteriormente. Para solventarlo, seguiremos la opción de marcar cada nivel en ficheros diferentes (la opción primera de las cuatro indicadas en la sección anterior).

Para tratar este problema, en trabajos anteriores (Navarro Colorado 2019) se propuso marcar en un mismo fichero tanto la información métrica como la categorial, y así se materializó en el *Corpus general de poesía lírica castellana del Siglo de Oro*.[28] Con este objetivo, se desarrolló un modelo de anotación basado en XML-TEI. En este modelo, las palabras quedan marcadas con la etiqueta correspondiente (“<w>”), tanto al inicio como al final de cada una, que contiene además información sobre el lema de la palabra y su categoría gramatical. La separación silábica, sin embargo, no se marca con la etiqueta “<seg>”, que es la recomendada por TEI, sino con la barra vertical (“|”). Este formalismo de la barra vertical es recomendado por las guías directrices de TEI para marcar la separación silábica de palabras en diccionarios. De esta manera, la sinalefa queda explícitamente marcada con una etiqueta fin de palabra (“</w>”) que no está precedida por la barra vertical indicativa de fin de sílaba. Véase la sinalefa producida en “para andar” en este verso de Lope de Vega:

```
<l met="-|-|-|-|5|-|7|-|" n="3">
   <w lemma="porque" type="CS">por|que|</w>
   <w lemma="para" type="SP">pa|ra</w>
   <w lemma="andar" type="VMN0000">an|dar|</w>
   <w lemma="conmigo" type="PP1CSO0">con|mi|go|</w>
</l>
```

28. https://github.com/bncolorado/CorpusGeneralPoesiaLiricaCastellanaDelSigloDeOro (19 de septiembre de 2023)

Actualmente, sin embargo, no considero apropiado este modelo. Al incluir además la información métrica en la etiqueta de verso <l>, la anotación resultante es muy densa y resulta compleja para la anotación manual. Por otro lado, su procesamiento automático no resulta óptimo al mezclar dos tipos de etiquetas, unas basadas en XML (la etiqueta de verso y palabra) y otra no (la barra vertical para separar sílabas).

En la propuesta actual considero que es mejor tratar cada fenómeno en ficheros diferentes. Con ello se busca una anotación más modular, de tal manera que se puedan marcar los textos en paralelo a diferentes niveles, y que en cualquier momento se puedan incorporar nuevos niveles de anotación sin que afecte a los niveles ya marcados.

Así, siguiendo las recomendaciones TEI y aprovechando la experiencia en el desarrollo de otros corpora literarios como ELTeC (Burnard *et al.* 2022), se parte de un fichero central que contiene el texto puro (la edición crítica digital) marcada únicamente con información textual básica: los metadatos propios del encabezado TEI y la estructural (títulos y versos). El resto de niveles de representación se marca en ficheros diferentes.

El siguiente código es un ejemplo de la anotación métrica. Como se puede ver, es una anotación XML-TEI estándar: las sílabas se marcan con la etiqueta genérica "<seg>" más el valor "syllable". El patrón métrico y rítmico con las etiquetas "<met>" y "<real>" respectivamente, más información sobre pausas y encabalgamientos. Para la justificación de este modelo de anotación, véase Navarro Colorado (2022):

```
<l source="#LopeVega.1063.1" met="--+--+-+-+-"
          real="--+--+-+-+-" enjamb="no">
<seg n="1" type="syllable" subtype="0"> Un</seg>
<w n="1"/>
<seg n="2" type="syllable" subtype="0"> so</seg>
<seg n="3" type="syllable" subtype="3"> ne</seg>
<seg n="4" type="syllable" subtype="0"> to</seg>
<w n="2"/> <pause type="1"/>
<seg n="5" type="syllable" subtype="0"> me</seg>
<w n="3"/>
<seg n="6" type="syllable" subtype="3"> man</seg>
<seg n="7" type="syllable" subtype="0"> da<w n="4"/> ha</seg>
<seg n="8" type="syllable" subtype="0"> cer</seg>
<w n="5"/>
<seg n="9" type="syllable" subtype="0"> Vio</seg>
<seg n="10" type="syllable" subtype="3"> lan</seg>
<seg n="11" type="syllable" subtype="0"> te</seg>
<w n="6"/>
<pause type="3"/>
</l>
```

El único dato ajeno a la métrica es el límite de las palabras. El final de cada palabra está marcado con la etiqueta <w>, pero como etiqueta simple <w/>. La razón por la que se incluye el límite de la palabra es para mantener la unidad del corpus: cada palabra (*token*) tiene un identificador único, que es utilizado en todos los niveles de anotación para poder relacionar unos ficheros con otros, independientemente del tipo de información anotada. Dado que estos identificadores actúan solo a modo de punteros, no es necesario hacer uso de la etiqueta de palabra "<w>" con apertura y cierre. Basta marcar el límite de cada palabra y su identificador con un elemento vacío"<w/>", lo que evita el problema de anidamiento. Seguimos en este punto la segunda recomendación de TEI para evitar este problema comentado anteriormente.

Otro fichero diferente contiene la anotación categorial del poema. El siguiente fragmento de código muestra un ejemplo. En este caso también se siguen las recomendaciones TEI. Esta anotación es léxica, no hay marcas para oraciones ni sintagmas, por lo que no hay problema de anidamiento. Cada palabra está localizada con su identificador único, que es el mismo utilizado en el nivel métrico y sintáctico:

```
<l source="#LopeVega.1063.1">
   <w n="1" lemma="uno" pos="DI0MS0"> Un</w>
   <w n="2" lemma="soneto" pos="NCMS000"> soneto</w>
   <w n="3" lemma="me" pos="PP1CS00"> me</w>
   <w n="4" lemma="mandar" pos="VMIP3S0"> manda</w>
   <w n="5" lemma="hacer" pos="VMN0000"> hacer</w>
   <w n="6" lemma="Violante" pos="NP00000">Violante</w>
</l>
```

Finalmente, en otro fichero diferente se están haciendo pruebas de anotación sintáctica con el modelo teórico de las gramáticas de dependencias y el formalismo CONLL (véase un ejemplo en la Tabla 1 mostrada anteriormente). El identificador de cada palabra es el mismo utilizado en el resto del corpus. La información morfológica y categorial en este caso es redundante, porque ya aparece en el fichero del nivel categorial. Por ahora no se elimina para mantener el formato estándar. Lo relevante es la información sintáctica: las relaciones de dependencia entre las palabras que se han explicado en la sección anterior.

Con esto ya se podría llevar a cabo el marcado necesario para el estudio del hipérbaton como rasgo estilístico y su relación con el ritmo poético. Se dispone por un lado de la posición de cada palabra en el verso, de su patrón métrico y rítmico, de la categoría gramatical de cada palabra y demás información morfológica, y de las relaciones sintácticas de dependencias entre las palabras.

En estos momentos está definido el modelo y se están desarrollando las pruebas de anotación. Cada nivel de anotación tiene ahora sus propios problemas que serán tratados en su momento, sobre todo la anotación sintáctica que, sin duda, es la más compleja. Con todo, lo que aquí hemos mostrado es cómo

integrar toda esta información, más otra que se pudiera añadir al corpus, en un modelo modular pero unificado, claro en la medida de lo posible, y que asegura la máxima compatibilidad con otros corpora y herramientas de PLN.

Como crítica a este modelo podría argumentarse, y con razón, que es un modelo redundante, pues hay información repetida en diferentes ficheros (sobre todo el propio texto). Esto no es apropiado desde un punto de vista computacional porque hace que sea necesario mucho espacio en disco para almacenar el corpus. Precisamente el modelo de marcado *stand-off* se diseñó para optimizar el espacio y evitar redundancias.

Por ahora se ha preferido mantener la redundancia para facilitar la anotación del corpus en paralelo por diferentes personas y con información diferente. Repetir el texto en cada fichero permite al anotador centrarse en la información que está marcando, sin depender de otros ficheros. Así, por ejemplo, anotar la sintaxis se puede realizar con independencia del fichero con información categorial. Finalizada la anotación del corpus se podrá plantear un modelo *stand-off* para su almacenamiento final. Gracias a los identificadores de palabra y al fichero central con el texto puro se podrá automatizar este paso; pero esto queda ya como trabajo futuro.

Conclusiones

En este capítulo se ha mostrado la necesidad, tanto computacional como literaria, de desarrollar corpora literarios anotados a diferentes niveles de descripción lingüístico-literaria. Para considerar un corpus así anotado como una edición digital enriquecida que sirva de referencia para los estudios literarios, se ha argumentado que el modelo debe ser compatible con otros corpora y herramientas computacionales; que así mismo debería relacionar los diferentes niveles de análisis para mantener la unidad textual, y, finalmente, que sería deseable que fuera un modelo lo suficientemente claro como para permitir la edición manual del código. Por último, se han presentado algunos problemas formales y los procedimientos para solventarlos, mostrando así la viabilidad de desarrollar un corpus anotado de poesía del Siglo de Oro que cumpla con estas características.

Bibliografía

Allés-Torrent, Susanna; Gabriel Calarco, y Gimena del Rio Riande, "Edición y publicación de textos con TEI", *TTHub. Text Technologies Hub: Recursos sobre tecnologías del texto y edición digital*, 2022 <https://tthub.io/aprende/tutorial/edicion-y-publicacion-textos-tei/>

Buchholz, Sabine y Erwin Marsi, "CoNLL-X shared task on Multilingual Dependency Parsing", *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*, New York City, 2006, pp. 149-164.

Burnard, Lou; Borja Navarro Colorado, Carolin Odebrecht y Martina Scholger, "Collaborative creation of a multi-lingual literary corpus. Challenges and best practices for corpus design", *COST Action Distant Reading Closing Conference*, Krakov/on-line, junio 2022.

Cabitza, F., Campagner, A., Basile, V., "Toward a Perspectivist Turn in Ground Truthing for Predictive Computing", *Proceedings of the AAAI Conference on Artificial Intelligence,* 2023 <https://arxiv.org/pdf/2109.04270.pdf>

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J. & Stede, M., "A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets", *Traitment automatique des langues* 49, (2008) pp. 271-293.

DeRose, Steven, "Markup Overlap: A Review and a Horse", *Proceedings of the Extreme Markup Languages*, Montréal, Canada, 2004.

Haider, Thomas, "Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features" *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3715-3725, 2021 <https://doi.org/10.18653/v1/2021.eacl-main.325>

Hatzel, Hans Ole; Haimo Stiemer; Chris Biemann y Evelyn Gius, "Machine learning in computational literary studies" *it - Information Technology* (2023) <https://doi.org/10.1515/itit-2023-0041>

Ide, Nancy y James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, Dordrecht, Springer, 2017.

Jauralde Pou, Pablo "Cerrar podrá mis ojos la postrera " *Revista de Filología Española*, LXXVII, 1/2 (1997).

Jurafsky, Dan y James H. Martin *Speech and Language Processing (3rd ed. draft)*, 2023 <https://web.stanford.edu/~jurafsky/slp3/>

Kesarwani, Vaibhav; Diana Inkpen; Stan Szpakowicz y Chris Tanasescu, "Metaphor Detection in a Poetry Corpus", *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Vancouver, Canada, 2017, pp. 1-9.

McShane, Marjorie y Sergei Nirenburg, *Linguistics for the Age of AI*, Cambridge, Massachusetts, MIT Press, 2021.

Meister , Jan Christoph, "Poetry, Phenomenon and Phenomenology" en Bories, Anne-Sophie; Petr Plecháč y Pablo Ruiz Fabo (eds.), *Computational Stylistics in Poetry, Prose, and Drama*, Berlin/Boston, de Gruyter, 2022.

Navarro Colorado, Borja; Ribes-Lafoz, María and Sánchez, Noelia "Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation", *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* Portorož, Eslovenia, 2016.

Navarro Colorado, Borja, "A Metrical Scansion System for Fixed-Metre Spanish Poetry", *Digital Scholarship in the Humanities* 33, 1 (2018*a*), pp. 112-127

Navarro Colorado, Borja, "On Poetic Topic Modeling: extracting themes and motifs from a corpus of Spanish poetry", *Frontiers in Digital Humanities. Computational Linguistics and Literature*, (2018*b*), DOI 10.3389/fdigh.2018.00015

Navarro Colorado, Borja "Por un análisis distante y profundo: un corpus piloto de la poesía lírica castellana del Siglo de Oro" *Revista de poética medieval*, 33 (2019).

Navarro Colorado, Borja, "Beyond the metre: formalization of linguistic-rhythmic features for a computational analysis of the Spanish poetry", *Calderón, R and Python: New Methods and Digital Tools for Quantitative Analysis of Theatre, Verse Drama and Poetry*, Vienna, University of Vienna, 2-3 junio 2022.

Piper, Andrew; Richard J. So y David Bamman, "Narrative Theory for Computational Narrative Understanding", *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 298-311. <https://doi.org/10.18653/v1/2021.emnlp-main.26>

Pustejovsky, James y Amber Stubbs, *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., 2012.

de la Rosa, Javier; Álvaro Pérez; Laura Hernández; Salvador Ros y Elena González-Blanco, "Rantanplan, Fast and Accurate Syllabification and Scansion of Spanish Poetry", *Procesamiento del lenguaje natural* 65 (2020), pp. 83-90

Schöch, Christof; Julia Dudar, Evegniia Fileva, eds., *Survey of Methods in Computational Literary Studies* Version 1.1.0, Trier, CLS INFRA, 2023 <https://methods.clsinfra.io>, DOI: 10.5281/zenodo.7892112.

The TEI Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange* 2023 <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

Torre, Esteban, "«Cerrar podrá mis ojos...»: Paráfrasis, métrica y hermenéutica" *Rhythmica*, 2, (2004), pp. 235 - 250 <https://doi.org/10.5944/rhythmica.6409>

Zeldes, Amir, "The GUM Corpus: Creating Multilayer Resources in the Classroom", *Language Resources and Evaluation* 51, 3 (2017), pp. 581-612.

Zipser, F. y L. Romary "A model oriented approach to the mapping of annotation formats using standards", *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*, Malta, 2010 <http://hal.archives-ouvertes.fr/inria-00527799/en/>