Journal of the Text Encoding Initiative
Rolling Issue | 2026

De Catulo a Wikidata

Automatización de tareas de codificación utilizando modelos de lenguaje, esquemas de metadatos y ontologías para un borrador de edición digital con el estándar XML-TEI

Carlos Javier Nusch, Gabriel Alejandro Calarco, Gimena del Rio Riande, Leticia Cecilia Cagnina, Leandro Antonelli y Marcelo Luis Errecalde

OpenEdition Journals

Edición electrónica
URL: https://journals.openedition.org/jtei/6987
ISSN: 2162-5603

Editor
TEI Consortium

Referencia electrónica
Carlos Javier Nusch, Gabriel Alejandro Calarco, Gimena del Rio Riande, Leticia Cecilia Cagnina, Leandro Antonelli and Marcelo Luis Errecalde, "De Catulo a Wikidata", *Journal of the Text Encoding Initiative* [Online], Rolling Issue, Online since 27 May 2026, connection on 27 May 2026. URL: http://journals.openedition.org/jtei/6987

De Catulo a Wikidata

Automatización de tareas de codificación utilizando modelos de lenguaje, esquemas de metadatos y ontologías para un borrador de edición digital con el estándar XML-TEI

Carlos Javier Nusch, Gabriel Alejandro Calarco, Gimena del Rio Riande, Leticia Cecilia Cagnina, Leandro Antonelli, y Marcelo Luis Errecalde

RESUMEN

Este artículo presenta un conjunto de procedimientos automatizados aplicados a la codificación y al análisis de un corpus poético que incluye las obras de Cayo Valerio Catulo, Albio Tibulo y Sexto Propercio. Para ello se diseñó un *pipeline* reproducible de procedimientos automatizados con el fin de codificar y analizar los textos latinos, integrando PLN con LatinCy (spaCy) y codificación XML-TEI. El flujo genera TEI con teiHeader y cuerpo, versos segmentados y numerados, marcado preliminar de entidades (personas, lugares o grupos) y anotación temática basada en el Diccionario de motivos amatorios de Moreno Soldevila mediante n-gramas y distancia de Levenshtein, implementada en tres modalidades TEI (*stand-off*, *flatten* e híbrida). Como productos principales, se obtuvieron 200 archivos TEI validados, un CSV consolidado de entidades con candidatos e identificadores recuperados desde VIAF, Pleiades y Wikidata (reutilizable para curaduría y

enriquecimiento posterior), y un conjunto de visualizaciones (barras y grafos de coocurrencia) para comparar patrones del imaginario amoroso entre autores; en la ejecución completa se registraron, además, 371 tópicos en Catulo, 450 en Tibulo y 730 en Propercio. Aunque los resultados no reemplazan la validación filológica (por ambigüedad, ruido de NER y falsos positivos/negativos en el *matching*), el enfoque ofrece una base técnica sólida para ediciones digitales semánticamente enriquecidas y para análisis exploratorios o cuantitativos con trazabilidad y supervisión editorial.

ABSTRACT

This article presents a set of automated procedures applied to the encoding and analysis of a poetic corpus comprising the works of Gaius Valerius Catullus, Albius Tibullus, and Sextus Propertius. To this end, we designed a reproducible pipeline of automated steps to encode and analyze Latin texts, integrating NLP with LatinCy (spaCy) and XML-TEI encoding. The workflow produces TEI documents with a teiHeader and body, segmented and numbered verses, preliminary tagging of entities (persons, places, or groups), and thematic annotation based on Moreno Soldevila's Dictionary of Amatory Motifs using character and word n-grams and Levenshtein distance, implemented in three TEI modalities (stand-off, flatten, and hybrid). The main outputs include 200 validated TEI files, a consolidated CSV of entities with candidates and persistent identifiers retrieved from VIAF, Pleiades, and Wikidata (reusable for subsequent curation and enrichment), and a set of visualizations (bar charts and co-occurrence graphs) to compare patterns in the amatory imaginary across authors; the full run additionally recorded 371 motifs in Catullus, 450 in Tibullus, and 730 in Propertius. Although these results do not replace philological validation (due to ambiguity, NER noise, and false positives/negatives in matching), the approach provides a solid technical basis for semantically enriched digital editions and for exploratory or quantitative analyses with traceability and editorial oversight.

ÍNDICE

Palabras Clave: edición digital, XML-TEI, poesía latina, PLN, LatinCy, reconocimiento de entidades nombradas, lematización, codificación temática, distancia de Levenshtein, visualización

Keywords: digital edition, XML-TEI, Latin poetry, NLP, LatinCy, Named Entity Recognition, lemmatization, thematic annotation, Levenshtein distance, visualization

DEDICATORIA

A Marcela Patronelli, mi profesora de literatura,
quien me honra con su amistad.
C. J. N.

NOTAS DEL AUTOR

Declaración de uso de IA generativa y tecnologías asistidas por IA en el proceso de escritura.
Durante la preparación de este trabajo, los autores utilizaron herramientas de IA generativa de OpenAI (incluyendo GPT-3.5, GPT-4 y GPT-5) para mejorar la legibilidad del manuscrito y realizar correcciones gramaticales y estilísticas menores. Además, se utilizaron herramientas de asistencia a la programación basadas en IA en el entorno de desarrollo (Codex de OpenAI en Visual Studio Code) para depurar errores, mejorar la claridad del código y facilitar tareas de refactorización y documentación de la versión publicada en GitHub. Tras utilizar estas herramientas, los autores revisaron y editaron los contenidos según fue necesario y asumen plena responsabilidad por el contenido final del artículo y del software descrito.

RECONOCIMIENTO

Agradezco a la Lic. Luciana Tanevitch por animarme y acompañarme durante el proceso de mejora del *pipeline*, y por su contribución clave para convertir el código en una versión más legible, ordenada y reproducible, adecuada para su publicación en GitHub

1. Introducción[1]

1 En trabajos anteriores hemos comentado las ideas de C. S. Lewis respecto a la influencia del amor cortés y la literatura occitana en el imaginario amoroso del siglo XX (Lewis 1936; Nusch et al. 2024, 2025). Se observó que había notables similitudes entre los temas amorosos, el tratamiento de la

persona amada y ciertos términos derivados de los campos político y militar en los poemas eróticos, tanto en la literatura occitana como en la latina del siglo I a.C. El objetivo principal que motiva estos trabajos es el intento de identificar patrones textuales en textos de la tradición literaria latina que pudieran arrojar luz sobre una posible herencia literaria de temas amorosos originada en la antigüedad y que culmina en el desarrollo del imaginario de la Religión del Amor por los poetas medievales occitanos. Se asumió un enfoque comparativo que intenta combinar técnicas tradicionales de lectura cercana con las capacidades de lectura distante proporcionadas por métodos computacionales (Jockers 2013; Moretti 2013; Ramsay 2011). El presente artículo examina el desempeño de distintos métodos de procesamiento del lenguaje natural en diversas tareas orientadas a automatizar la codificación y la estructuración de archivos conforme al estándar XML-TEI, algunas de las cuales ya habían sido propuestas previamente en la tesis de maestría (Nusch 2021).

2. Definición del Problema, Objetivos y Contribuciones

2 En los trabajos referidos anteriormente se propusieron diferentes posibilidades para el abordaje de la temática amorosa en la Antigüedad y la Edad Media por medio de la codificación de textos. Se hizo un primer análisis de posibles etiquetas, bases de datos de autoridades, identificadores persistentes y modos de marcar los distintos tópicos amorosos en los poemas escogidos. También se exploraron técnicas de agrupamiento de textos como el algoritmo K-means (Lloyd 1982) o la extracción de términos típicos de la poesía amorosa con árboles de decisión (Quinlan 1986).

3 En esta ocasión presentamos varias tareas diseñadas para automatizar la codificación de los poemas de nuestro corpus con el estándar XML-TEI. Se diseñó un *pipeline* que contemplaba diferentes etapas y que buscaba evitar tareas tediosas y repetitivas a los editores (marcado de datos comunes a todos los poemas como la información de edición o el conteo y marcado de número de versos). Además, se realizó, por medio de técnicas de minería de textos, el reconocimientopotencial de el reconocimiento y premarcado de entidades nombradas[2] y la recuperación de tópicos amorosos que requerirá, por supuesto, un posterior control y supervisión de especialistas. También se exploraron diferentes bases de datos de autoridades con el objeto de obtener identificadores

persistentes que luego pudieran usarse, en una etapa posterior de este proyecto, para enriquecer los datos de personas, grupos y lugares de esta edición y hacerla interoperable. El *pipeline* se divide en las siguientes etapas:

1. Codificación automática del `<teiHeader>`.
2. Conteo automático de versos y posterior codificación.
3. Reconocimiento de entidades nombradas.
4. Codificación automática de entidades en XML-TEI.
5. Consulta de bases de datos de autoridades y búsqueda de identificadores persistentes (VIAF,[3] Pleiades,[4] Wikidata[5]) y guardado de los datos en un archivo CSV para el asignado manual en una etapa futura del proyecto.
6. Recuperación automatizada de motivos amorosos en Catulo, Tibulo y Propercio utilizando como recurso esencial y criterio taxonómico el *Diccionario de motivos amatorios en la literatura latina (siglos III a. C. -II d. C)* (Moreno Soldevila 2011).
7. Codificación automática de los tópicos recuperados en textos XML-TEI.
8. Validación en masa del esquema Relax NG.
9. Generación de imágenes y visualizaciones posteriores para la lectura distante.

3. Estado del arte

4 El problema del reconocimiento de entidades nombradas para el latín ya fue tratado previamente por Erdmann et al. (2016) quienes desarrollaron el primer corpus anotado en latín para este tipo de tarea, utilizando una estrategia de aprendizaje activo y un modelo supervisado entrenado sobre textos del corpus Perseus. Su enfoque permitió alcanzar una efectividad de más del 90 %[6] en pruebas dentro del dominio, superando ampliamente a herramientas previas. Su estudio destacó los desafíos específicos del latín, como la escasez de datos anotados, la variación estilística de los textos y la complejidad morfológica propia de una lengua flexiva. El mismo año Torres Aguilar et al. (2016) presentaron un modelo de reconocimiento de entidades nombradas (NER) aplicado a un corpus de 5300 documentos en latín medieval, provenientes de cartularios de abadías cluniacenses y cistercienses de Borgoña de los siglos X y XIII. El corpus, anotado manualmente con información

léxica, morfológica y semántica, se utilizó para entrenar un modelo basado en Campos Aleatorios Condicionales (CRF)[7] con el objetivo de automatizar la identificación de nombres de personas, lugares e instituciones.

5 Más recientemente, David Bamman y Patrick J. Burns (2020) presentaron el desarrollo de Latin BERT, un modelo de lenguaje contextual entrenado con más de 640 millones de palabras provenientes de diversas fuentes en latín (Corpus Thomisticum, Internet Archive, Latin Library, Patrologia Latina, Perseus y Latin Wikipedia). El modelo alcanzó un nuevo estado del arte en tareas como el etiquetado morfosintáctico, la desambiguación léxica y la predicción de texto faltante. Entre sus posibles usos también se encuentra el reconocimiento de entidades nombradas. Años después, Burns (2023) presentó LatinCy, un conjunto de modelos entrenados para el procesamiento de lenguaje natural en latín dentro del marco spaCy[8] construido a partir de la integración de cinco árboles de dependencia universal y otros corpus. Los modelos lograron altos niveles de precisión en tareas como el etiquetado morfosintáctico (97.4%), la lematización (94.7%) y el reconocimiento de entidades (hasta 90.8%).[9]

6 Con respecto a la codificación elegida para este trabajo, se pueden citar artículos anteriores que reflexionan sobre el tema, como Piotr Bański (2010), quien, en su análisis sobre la codificación *stand-off*, destaca su sostenibilidad, modularidad y flexibilidad, ya que permite preservar el texto fuente sin alteraciones y aplicar múltiples capas de codificación independientes. Sin embargo, también advierte que su implementación práctica enfrenta obstáculos importantes, como la falta de soporte para determinados estándares (XInclude/XPointer). Raffaele Viglianti (2016) igualmente notó que, pese a su potencial teórico, la adopción de la codificación *stand-off* en TEI ha sido limitada debido a la complejidad de su implementación. En proyectos que requieren representaciones complejas o múltiples capas de codificación, este tipo de codificación con modelos de hitos suele ser difícil de mantener y la introducción manual de identificadores es propensa a errores. Para resolver estos desafíos, el autor propuso simplificar la codificación mediante herramientas específicas como coreBuilder, una aplicación web que permite crear codificaciones *stand-off* en un entorno visual, que reduce errores y mejora la eficiencia.

7 En el ámbito de la georreferenciación y la vinculación con entidades e identificadores persistentes, un trabajo mucho más amplio y abarcativo que el que se propone en este artículo es el de Ciotti et al. (2014), quienes desarrollaron un enfoque que combina la codificación XML-TEI con

modelos ontológicos y prácticas de datos abiertos vinculados para enriquecer la representación semántica de textos latinos. A través del proyecto Geolat,[10] los autores demostraron cómo la codificación TEI puede ser utilizada de forma flexible para anotar textos con múltiples niveles de profundidad, y cómo la vinculación con ontologías (RDF) permite formalizar el conocimiento literario y geográfico, para una mejor interoperabilidad con otros recursos y bases de datos (Pleiades, Pelagios, Wikipedia, VIAF).

8 Con respecto a la recuperación automatizada de motivos o tópicos amatorios es necesario aclarar que en esta ocasión no se realizó una tarea de modelado de tópicos tradicional al estilo de las realizadas anteriormente con técnicas como *Latent Dirichlet Allocation* (LDA) (Nusch 2021) sino que se llevó a cabo de otro tipo de tarea de minería de textos, más específicamente de extracción o recuperación de la información por medio del uso de la distancia de Levenshtein (1965) utilizando como recurso principal el *Diccionario de motivos amatorios en la literatura latina (siglos III a. C. - II d. C.)* coordinado por Rosario Moreno Soldevila (2011). Esta obra es el trabajo especializado que posibilitó la tarea de minería de textos ya que presenta más de 160 motivos amatorios propios de la literatura latina organizados alfabéticamente (temas como *adulterio*, *amada*, *amado*, *amor*, *belleza*, *besos*, *coito*, *Cupido*, *dioses del amor*, *esclavitud de amor*, *lesbianismo*, *milicia de amor*, *sexo*, *síntomas de amor*, *Venus*, etc.). Por medio del sistema desarrollado en este trabajo y siguiendo la clasificación taxonómica propuesta en el diccionario, se ha conseguido aprovechar el conocimiento y las décadas de trabajo de los mejores especialistas en la materia y traerlos, con la ayuda de técnicas de procesamiento del lenguaje natural, a cumplir el rol de codificadores expertos virtuales en nuestra futura edición digital.

4. Metodología de investigación y enfoque

4.1 Corpus de análisis y ediciones utilizadas

9 El corpus de trabajo en esta ocasión (tabla 1) incluye las obras completas de Cayo Valerio Catulo (Merrill 1893), Albio Tibulo (Postgate 1915) y Sexto Propercio (Müller 1898).

Tabla 1. Tamaño del corpus de estudio en versos, palabras totales (*tokens)*, palabras únicas (*types*) y promedio de palabras por poema o canto. Elaboración propia.

Author / Work	Verses	Total Words	Unique Words	Average Words per Poem/Canto
Catulo (Merrill, 1893)	2289	12912	5802	110.35
Tibulo (Postgate, 1915)	1930	12368	5201	334.27
Propercio (Müller, 1898)	4008	25450	9809	242.38

4.2 Enfoque general del flujo de trabajo

10 El objetivo de este trabajo no es solamente producir una edición anotada, sino documentar un flujo de trabajo reproducible que combine herramientas de PLN, esquemas XML-TEI y el potencial del reconocimiento de entidades nombradas para obtener recursos de datos enlazados (VIAF, Pleiades, Wikidata, entre otros) y también de determinadas técnicas de minería de textos, como el uso de la distancia de Levenshtein en la identificación de tópicos amorosos a la hora de asistir a los editores que, en todo momento, deben supervisar la edición digital final.

11 El proceso se organiza en una serie de módulos encadenados que pueden ejecutarse de forma independiente o como parte de un pipeline completo, como puede observarse en las figuras 1 y 2. Los diferentes procesos se organizan de la siguiente manera:

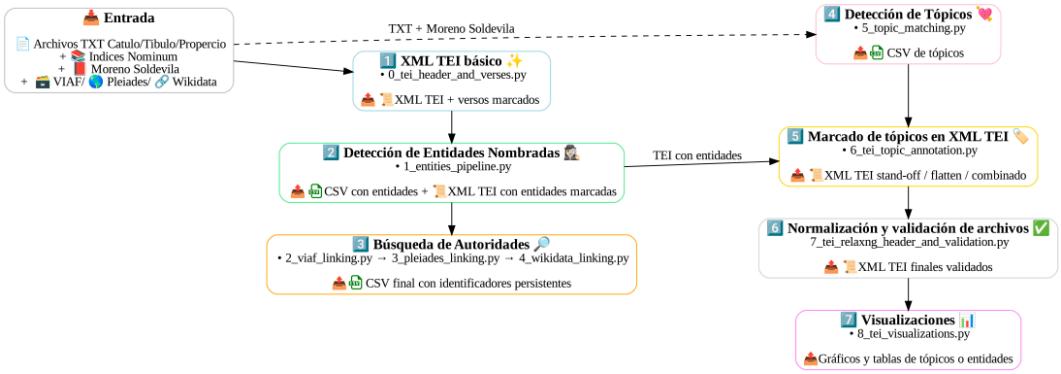
1. Preparación del texto a partir de archivos TXT y generación del `<teiHeader>` inicial, con la descripción bibliográfica de las ediciones utilizadas, la información sobre el proyecto y la taxonomía utilizada para los tópicos amorosos.[11]
2. Identificación y numeración de versos, con transformación del texto original a una estructura XML-TEI basada en `<lg>` (grupos de versos) y `<l>` (versos individuales).
3. Reconocimiento automático de entidades nombradas (personas, lugares y grupos) en los versos mediante un modelo de lenguaje entrenado para latín (LatinCy).

    3.1 Anotación automática de entidades en los archivos XML-TEI.

3.2 Normalización y lematización de las formas reconocidas, y construcción de tablas intermedias con información sobre entidad, lema, tipo y contexto.

4. Búsqueda de las entidades mediante consultas a APIs de VIAF, Pleiades y Wikidata, aplicando filtros por tipo de entidad y cronología[12] y generación de un archivo CSV con los datos de las diferentes entidades e identificadores persistentes cosechados (estos datos se incluirán en un trabajo posterior que aún no se ha realizado en el que evaluaremos exhaustivamente los resultados obtenidos y realizaremos un marcado manual de los diferentes identificadores).
5. Recuperación automática de motivos o tópicos amorosos a partir del diccionario especializado y generación de anotaciones temáticas vinculadas a fragmentos de verso. Anotación de los diferentes tópicos amorosos recuperados en los poemas en formato XML-TEI.
6. Generación de diferentes variantes de marcado TEI: anotaciones temáticas en *stand-off*, *flatten* y combinaciones de ambas.
7. Validación de los resultados mediante esquemas Relax NG.
8. Producción de visualizaciones (grafos de entidades, mapas de coocurrencias de tópicos, etc.) que permiten explorar los resultados.

Figura 1. Vista general del workflow desarrollado con diferentes módulos, archivos de entrada, operaciones intermedias y archivos de salida. Elaboración propia.



12 En los apartados que siguen se describen con más detalle los modelos y recursos utilizados, así como los criterios de filtrado y de aceptación de la información obtenida automáticamente de las bases de datos de autoridades y datos enlazados.

4.3 Modelos, recursos y herramientas utilizadas

13 Para las tareas de procesamiento del lenguaje natural sobre el corpus se utilizó LatinCy,[13] basado en spaCy, que ofrece tokenización, lematización y reconocimiento de entidades nombradas adaptados al latín. Este modelo se aplicó para reconocer entidades en los poemas y generar dos salidas complementarias: (i) los archivos XML-TEI enriquecidos con marcado de entidades y (ii) una tabla intermedia en formato CSV que registra cada ocurrencia junto con un contexto editorial mínimo (p. ej., autor/obra/poema). Este insumo tabular se reutiliza luego para el enriquecimiento con identificadores persistentes, manteniendo la trazabilidad entre el texto editado y los resultados de consulta a fuentes externas. En esta etapa, el objetivo no fue completar de manera automática y definitiva la identificación correcta de cada entidad, sino automatizar tareas repetitivas y producir sugerencias estructuradas que puedan ser revisadas por especialistas en fases posteriores.

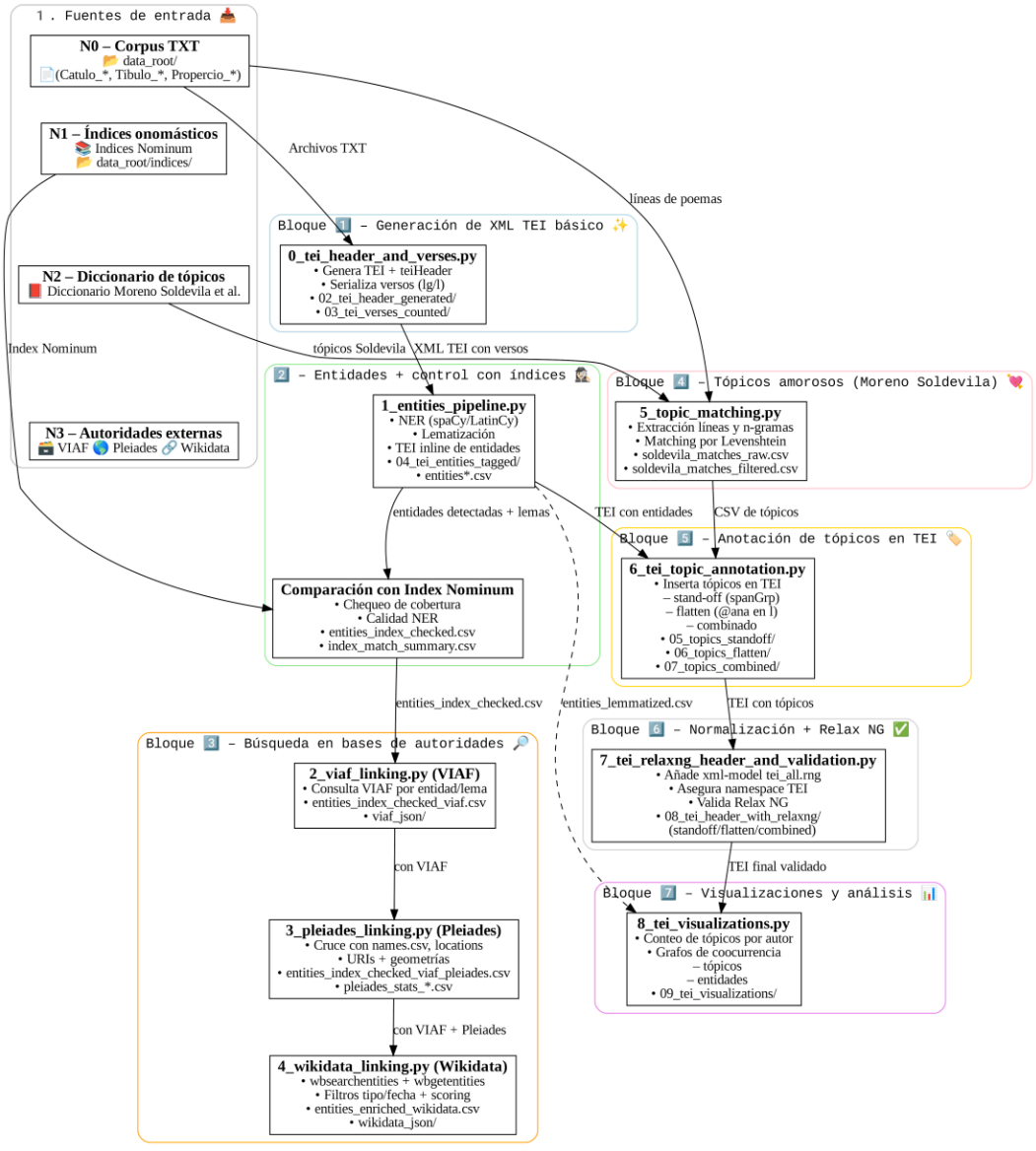
14 A partir de las ocurrencias detectadas en Catulo, Tibulo y Propercio se construyó un inventario de 2921 pares entidad–lema (1727 personas, 813 lugares y 381 grupos o colectivos), que constituye la base del proceso de enriquecimiento. Sobre ese inventario se ejecutaron consultas automáticas a tres recursos complementarios: VIAF (autoridades normalizadas, especialmente para personas), Pleiades (lugares del mundo antiguo con información geográfica) y Wikidata (grafo de conocimiento para personas, lugares y colectivos). Para maximizar la cobertura, el sistema conserva múltiples candidatos por cada entidad consultada, operando tanto sobre la forma flexionada tal como aparece en el texto (*entity*) como sobre el lema normalizado (*lemma*). En el estado actual, el volumen de candidatos devueltos por las APIs alcanza decenas de miles de combinaciones entidad–candidato, lo que justifica separar la obtención automática de candidatos de la selección definitiva, que se concibe como una tarea filológica, manual y evaluable en una etapa posterior.

4.4 Criterios de filtrado de resultados

15 Para maximizar cobertura sin perder trazabilidad, en esta fase se aplica un filtrado intencionalmente conservador: en lugar de decidir un identificador único "correcto" por mención, se prioriza construir un conjunto amplio de candidatos plausibles que el editor pueda revisar, corregir y refinar en etapas posteriores. En consecuencia, el sistema preserva alternativas y registra explícitamente el origen de cada candidato (por ejemplo, si proviene de la forma flexionada o del lema), de modo que la validación filológica posterior pueda auditar tanto los aciertos como los falsos positivos.

16 Los criterios operativos concretos —incluyendo la organización de resultados por consultas sobre *entity*/*lemma*, la distinción entre variantes estrictas y laxas, y la separación de candidatos con/sin información temporal cuando corresponde— se detallan en sección 5.3.3 en el marco de las consultas a Wikidata, donde además se presenta el esquema de priorización y la forma en que se conservan los candidatos mejor posicionados para análisis y revisión editorial. En las secciones metodológicas generales, por tanto, se mantiene únicamente este encuadre: la etapa automatizada se orienta a maximizar recuperación y trazabilidad; la selección definitiva de identificadores persistentes queda deliberadamente diferida a una fase posterior de evaluación experta.

Figura 2. Vista detallada del workflow desarrollado con diferentes módulos, ubicaciones, archivos de entrada, operaciones intermedias y archivos de salida. Elaboración propia.



5. Paso a paso: tareas de PLN y marcado XML TEI realizadas

17 A continuación se describe en detalle cada tarea realizada en el orden en el que se ejecuta cada parte del workflow.

5.1 Una codificación de cinco minutos: creación del <teiHeader>, conteo de versos y marcado

18 La tarea inicial consistió en crear documentos en formato TEI a partir de un conjunto de archivos de texto plano correspondientes a poemas de *Catulo*, *Tibulo* y *Propercio*, previamente obtenidos de la Biblioteca Digital Perseus[14] (Cerrato y Chavez s.f.). El objetivo de esta primera etapa fue construir una estructura básica conforme al estándar XML-TEI, incorporando los elementos mínimos requeridos. Para ello se desarrolló un procedimiento automatizado en Python que recorre sistemáticamente cada archivo `.txt` y genera un archivo XML estructurado, compuesto por los elementos principales `<TEI>`, `<teiHeader>`, `<text>` y `<body>`.

19 El encabezado `<teiHeader>` se generó dinámicamente a partir del nombre del poema y del autor. Al leer el nombre de la carpeta que contiene el archivo, el sistema infiere el nombre del autor (por ejemplo, *Catulo*) y lo transforma en su forma completa en latín (*Gaius Valerius Catullus*). De la misma forma, se extrajo el título del poema desde el nombre del archivo normalizando su formato para construir un identificador único (`@xml:id`). En la sección `<titleStmt>` se incorporan, además, los nombres de los responsables de la edición digital, codificados mediante elementos `<respStmt>` con la designación de `"Edición digital"` como función.

20 La sección `<sourceDesc>` del encabezado contiene una referencia bibliográfica a la edición impresa utilizada como fuente. Esta información se representa mediante el elemento `<bibl>`, en un formato estructurado equivalente al estilo APA: nombre del editor, año de publicación, título de la edición, editorial y un enlace al ejemplar digital. Complementariamente, dentro de `<encodingDesc>`, se documenta el modelo de procesamiento de lenguaje natural utilizado para el análisis automático por medio de los atributos `@version="3.8.0"` e `@ident="latincy_la_core_web_lg"`.[15]

21 Además de la estructura bibliográfica y técnica del encabezado, en esta misma etapa se incorporó el recurso conceptual que organiza la anotación temática del proyecto: el *Diccionario de motivos amatorios en la literatura latina* de Moreno Soldevila (2011). Para ello se declara una taxonomía TEI en `<encodingDesc>/<classDecl>` que funciona como vocabulario controlado de motivos. Esta declaración cumple dos funciones: por un lado, documenta explícitamente la fuente intelectual de la clasificación temática; por otro, establece un identificador estable (`@xml:id`) que luego en la etapa de marcado de tópicos permite referenciar los motivos al anotar pasajes del texto (por ejemplo, mediante `@ana` o `@corresp`) sin ambigüedad ni duplicación de etiquetas.

22 Cada poema resultante se serializa como un archivo XML válido que incluye tanto el encabezado enriquecido como el cuerpo textual dentro de `<text>/<body>`, inicialmente representado como un único bloque plano de texto.

23 Una segunda etapa del procesamiento reemplazó esta estructura plana por una codificación más precisa del verso, transformando el contenido del cuerpo en una secuencia de elementos `<l>` numerados mediante el atributo `@n`, contenidos dentro de un bloque `<lg>`. Esta transformación, sencilla desde el punto de vista computacional, permite una representación estructurada de la métrica y habilita análisis métricos, segmentaciones y visualizaciones avanzadas como puede verse en el ejemplo 1.

24 En etapas posteriores, todos los archivos generados fueron procesados para asegurar su conformidad con el esquema formal del TEI. Esto implicó eliminar declaraciones XML redundantes e incorporar una directiva xml-model al inicio del archivo, que referencia el esquema RelaxNG oficial del TEI (tei_all.rng). Asimismo, se garantizó la presencia del atributo de espacio de nombres `xmlns="http://www.tei-c.org/ns/1.0"` en el nodo raíz `<TEI>` cuando fuera necesario.

25 Finalmente, todos los archivos fueron validados automáticamente contra el esquema TEI P5 en formato RelaxNG, cargado directamente desde el sitio oficial de la Text Encoding Initiative. Los errores de sintaxis o validación estructural detectados durante este proceso fueron registrados en un archivo de texto para su revisión posterior.

Ejemplo 1. Vista de la primera versión del archivo TEI con el header y body codificados automáticamente. Elaboración propia.

```
<TEI>
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title xml:id="carmen005">Carmen 005</title>
    <author>Gaius Valerius Catullus</author>
    <respStmt>
     <name>Carlos Nusch</name>
     <resp>Edicion digital</resp>
    </respStmt>
    <respStmt>
     <name>Gabriel Calarco</name>
     <resp>Edicion digital</resp>
```

```
    </respStmt>
    <respStmt>
     <name>Gimena del Rio Riande</name>
     <resp>Edicion digital</resp>
    </respStmt>
   </titleStmt>
   <publicationStmt>
    <publisher>Proyecto Aetatis Amoris</publisher>
    <availability>
     <licence>
      <ref target="https://creativecommons.org/licenses/by/4.0/">CC BY 4.0</ref>
     </licence>
    </availability>
   </publicationStmt>
   <sourceDesc>
    <bibl>
     <author>Merrill, E. T.</author>
     <date>1893</date>
     <title>Catullus; edited by Elmer Truesdell Merrill</title>
     <publisher>Boston Ginn</publisher>
     <ptr target="http://archive.org/details/catulluseditedby00catuuoft"/>
    </bibl>
   </sourceDesc>
  </fileDesc>
  <encodingDesc>
   <appInfo>
    <application version="3.8.0" ident="latincy_la_core_web_lg">
     <label>la_core_web_lg v3.8.0</label>
     <ptr target="https://huggingface.co/latincy/la_core_web_lg"/>
    </application>
   </appInfo>
   <classDecl>
    <taxonomy xml:id="soldevila">
     <bibl>
      <author>Moreno Soldevila, R. (Ed.)</author>
      <date>2011</date>
      <title>Diccionario de motivos amatorios en la Literatura Latina</title>
      <publisher>Universidad de Huelva</publisher>
```

```
      </bibl>
     </taxonomy>
    </classDecl>
   </encodingDesc>
  </teiHeader>
  <text>
   <body>
    <lg type="poema">
     <l n="1">Vivamus, mea Lesbia, atque amemus,</l>
     <l n="2">rumoresque senum severiorum</l>
     <l n="3">omnes unius aestimemus assis.</l>
     <l n="4">soles occidere et redire possunt:</l>
     <l n="5">nobis, cum semel occidit brevis lux,</l>
     <l n="6">nox est perpetua una dormienda.</l>
     <l n="7">da mi basia mille, deinde centum,</l>
     <l n="8">dein mille altera, dein secunda centum,</l>
     <l n="9">deinde usque altera mille, deinde centum,</l>
     <l n="10">dein, cum milia multa fecerimus,</l>
     <l n="11">conturbabimus illa, ne sciamus,</l>
     <l n="12">aut ne quis malus invidere possit,</l>
     <l n="13">cum tantum sciat esse basiorum.</l>
    </lg>
   </body>
  </text>
</TEI>
```

5.2 Reconocimiento de Entidades Nombradas

26 En el siguiente paso se realizó un proceso de reconocimiento de entidades nombradas, método que consiste en identificar y clasificar automáticamente menciones de entidades específicas dentro de un texto (personas, lugares, organizaciones, grupos, etcétera). En el análisis de textos latinos, como en este caso, dicha tarea permite reconocer nombres propios que remiten a personajes, topónimos o grupos sociales.

27 Como se indicó anteriormente, empleó el modelo la_core_web_lg_de la biblioteca spaCy, una de las variantes del proyecto LatinCy (Burns 2023) entrenadas específicamente para latín. Se utilizó un enfoque zero-shot para el procesamiento y, para cada archivo, se extrajeron todas las

entidades reconocidas por el modelo, almacenando su nombre, tipo y archivo de origen. *Zero-shot* (o "aprendizaje sin ejemplos") se refiere a un enfoque en el que un modelo realiza una tarea —por ejemplo, clasificar, etiquetar o relacionar información— sin haber sido entrenado específicamente con ejemplos anotados para ese conjunto de datos o para ese corpus en particular; en su lugar, se apoya en conocimiento previo adquirido durante su entrenamiento general y en la formulación de la consigna o contexto provisto. En nuestro caso, el reconocimiento de entidades nombradas se aplica directamente sobre el corpus de poemas con LatinCy, sin entrenar ni ajustar el modelo previamente con anotaciones propias del proyecto; el modelo produce una primera propuesta automática de entidades que luego debe ser revisada y validada por los editores.

28 Los tipos de entidades reconocidas en este estudio fueron:

- *PERSON*: personas o personajes individuales;
- *LOC*: lugares geográficos;
- *NORP*: colectivos, pueblos, grupos nacionales o religiosos.

29 Los resultados fueron los siguientes: en el corpus de Catulo se identificaron 427 personas, 185 lugares y 59 grupos; en Tibulo, 391 personas, 144 lugares y 88 grupos; y en Propercio, 909 personas, 484 lugares y 234 grupos. El recuento ofreció una primera aproximación cuantitativa al universo de entidades en cada corpus poético. Este procedimiento también permitió evaluar el rendimiento del modelo en el corpus y se pudieron identificar diferentes tipos de errores en el reconocimiento automático de entidades:

1. En algunos casos, la palabra identificada no es una entidad sino alguna palabra que logró confundir al modelo (por el uso de mayúsculas o por estar al inicio de oración).
2. En otros casos la palabra identificada sí es una entidad pero no el tipo de entidad reconocida (por ejemplo, Lesbia, la amada de Catulo apareció identificada como lugar cuando debió identificarse como persona)
3. Un último caso que se identificó es el siguiente, aunque no se trata propiamente de un error: en muchas ocasiones se suele hacer un uso metafórico de personajes mitológicos para referirse a lugares (Urano, el dios del cielo para referirse al cielo o Scilla y Caribdis, que además de ser monstruos mitológicos, funcionan como referencias geográficas al estrecho de Mesina).

30 Con respecto al primer tipo de error, se debe aclarar que el modelo actual de LatinCy ha mejorado notablemente respecto al de 2024 y ya no reconoce como entidades términos que no lo son. El segundo tipo de error, la identificación correcta de una entidad pero no su tipo, puede explicarse por la cercanía de los significantes, es decir, de los caracteres entre nombres de personas, Delia, Lesbia, Galo, por ejemplo con nombres de lugares como la isla de Lesbos o la región de la Galia o la isla de Delos (puede verse el resultado de este tipo de error en la entidad anotada automáticamente *Lesbia* en el ejemplo 2). Otro caso que puede facilitar la confusión es cuando una región o lugar se nombra en honor a una persona, algo muy común en todas las épocas y también en la Antigüedad (Roma por Rómulo, Troya por Tros, Alejandría por Alejandro, etcétera). Estas particularidades que pueden confundir a los modelos deben ser salvadas en una tarea posterior de control de resultados por parte de los editores y filólogos.

31 A partir de las entidades reconocidas automáticamente en el corpus se construyó en un archivo CSV un inventario de 2921 pares entidad–lema, compuesto por 1727 PERSON, 813 LOC y 381 NORP. Esta tabla intermedia conserva, como mínimo, la forma tal como aparece en el verso, su lema normalizado, el tipo de entidad y la referencia editorial (obra y poema de aparición). Este diseño permite separar con claridad dos objetivos: por un lado, la anotación preliminar en el TEI; por otro, la construcción de un insumo trazable para enriquecer manualmente la edición con identificadores persistentes y metadatos normalizados.

5.2.1 Bases de datos de autoridades consultadas

32 En un primer momento, y como se señaló anteriormente (Nusch 2021), se optó por utilizar la base de datos de autoridades VIAF como fuente principal para asignar a las entidades un identificador persistente en la edición digital. Sin embargo, en el transcurso del trabajo se evidenciaron ciertas limitaciones: VIAF, por ejemplo, no posee un registro detallado de personajes mitológicos como los que se suelen encontrar en los poetas estudiados y en lo que respecta al filtrado de entidades por fecha, al alimentarse de diferentes fuentes (Library of Congress, Deutsche Nationalbibliothek, Bibliothèque nationale de France, entre muchísimas otras) los datos expuestos no siguen un patrón o normalización única fácilmente minable y reutilizable. Por este motivo, se decidió complementar el uso de VIAF con Wikidata, cuya estructura basada en propiedades permite una normalización más precisa de fechas (por ejemplo, nacimiento, muerte o período de actividad) y una mayor flexibilidad en la vinculación de datos. A su vez, Wikidata ofrece la posibilidad de recuperar,

junto con el identificador persistente, información adicional de relevancia, como representaciones visuales, mapas geográficos, relaciones genealógicas o vínculos con otras bases de conocimiento, lo que la convierte en una herramienta óptima para la codificación enriquecida de entidades como la que se planificó en el proyecto. Como tarea adicional para los topónimos vinculados a la geografía del mundo antiguo, se incorporó la base de datos Pleiades, ampliamente utilizada en los estudios clásicos y proyectos de humanidades digitales.

5.2.2 Integración con ediciones digitales en XML-TEI

33 La información recuperada a través de Wikidata tiene múltiples aplicaciones dentro de una edición digital basada en TEI. En primer lugar, permite añadir atributos como `@ref`, `@type`, `@source` y `@when` a las marcas de entidades `<persName>`, `<placeName>`, vinculándolas de forma explícita con datos estructurados y abiertos. Esta vinculación incrementa la interoperabilidad del corpus con otras ediciones o proyectos de linked open data (LOD) en humanidades digitales.

34 En segundo lugar, el enriquecimiento semántico con Wikidata puede posibilitar la incorporación futura de líneas cronológicas, mapas históricos, perfiles biográficos o visualizaciones interactivas, todo ello basándose en los datos descargados y alineados con los estándares TEI. Por ejemplo, una persona mencionada en los textos puede estar vinculada a un `<persName>` con una referencia directa a su entrada en Wikidata, complementada con información cronológica, imágenes u otras relaciones semánticas relevantes.

5.2.3 Lematización y control de entidades

35 Otra técnica de procesamiento del lenguaje natural utilizada fue la lematización, que consiste en el proceso de reducir una palabra a su forma base o canónica, conocida como lema. El latín es una lengua altamente flexiva y las palabras pueden presentar numerosas variaciones morfológicas según el caso, número, género, tiempo, modo o persona. Este proceso permite agrupar todas esas formas bajo una misma entrada léxica y facilitar su búsqueda en bases de datos. El modelo de LatinCy utilizado incorpora herramientas de análisis morfosintáctico y lematización entrenadas específicamente para el latín. La lematización es especialmente útil al tratar con entidades nombradas, ya que en el caso del latín los nombres propios tienden a almacenarse en bases de datos en su forma de nominativo singular (por ejemplo, *Aeneas*), mientras que en los textos pueden aparecer en otras formas, como *Aenean* (acusativo, “a Eneas”) o *Aeneae* (genitivo, “de Eneas”).

Normalizar todas estas variantes en un mismo lema permite mejorar la comparación con recursos externos (como VIAF o Wikidata), garantizar una mejor visualización de grafos de coocurrencia y evitar duplicaciones en los análisis de frecuencia. No obstante, las búsquedas en las bases de datos se realizaron tanto en forma lematizada como sin lematizar para aumentar la posibilidad de obtener resultados y ante la posibilidad de que la lematización, un proceso automático, no haya sido correcta.

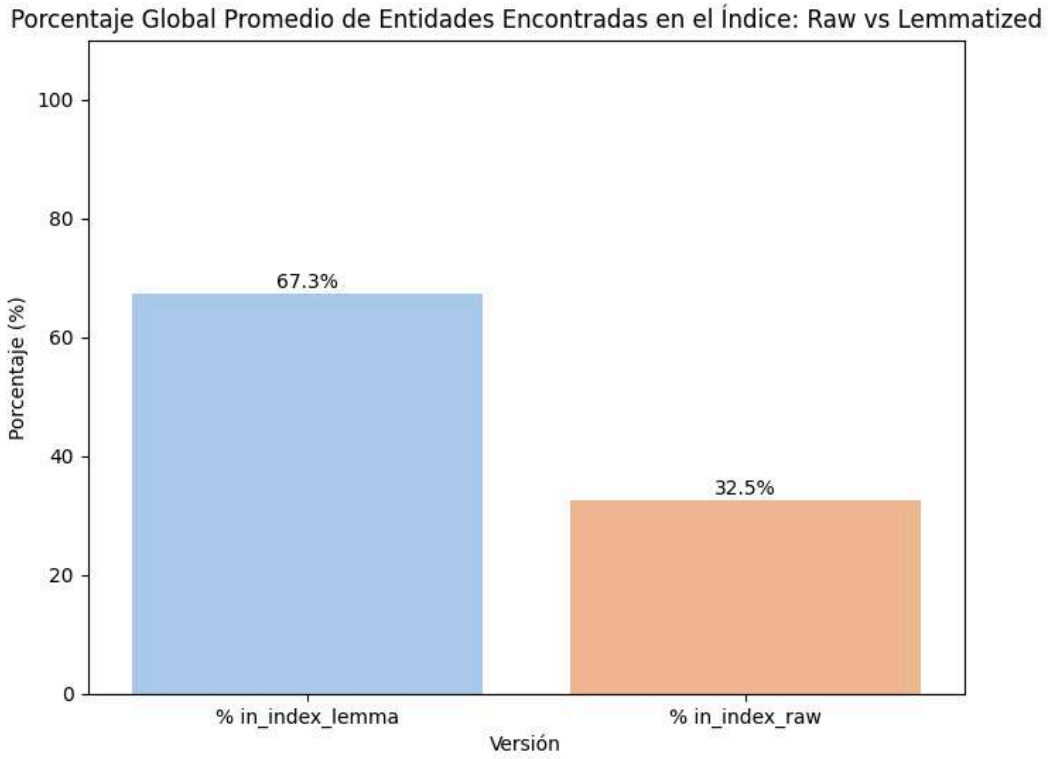
5.2.4 Comparación con *indices nominum*

36 Otra tarea de validación de datos que se llevó a cabo fue evaluar provisoriamente la fiabilidad de las entidades reconocidas por el modelo LatinCy: se realizó un análisis comparativo entre las entidades reconocidas (tanto en su forma original como lematizada) y los nombres propios registrados en los *indices nominum* de las ediciones críticas utilizadas en este proyecto.[16] Este cotejo permitió identificar posibles errores de detección, especialmente en entidades de tipo persona (PERSON) y lugar (LOC), las cuales suelen estar bien documentadas en dichos repertorios onomásticos. Para ello, se procesaron los poemas de Catulo, Tibulo y Propercio, extrayendo ambas variantes (palabra original y lema) de cada entidad y verificando su presencia en los respectivos índices.

37 La comparación con los *indices nominum* reveló una cantidad no menor de entidades —mayormente lematizadas— que no figuran en dichos repertorios. En el caso de Catulo, el 30,21% de las entidades PERSON lematizadas no se hallan en el índice correspondiente, porcentaje que asciende al 33,76% en Propercio y al 33,75% en Tibulo. Esta divergencia podría obedecer a múltiples factores: errores de segmentación o etiquetado por parte del modelo, variantes formales no reconocidas, o bien menciones contextuales que escapan al alcance de los repertorios tradicionales. Ante esta situación, será imprescindible introducir una fase de supervisión manual y filológica en un trabajo futuro, que evalúe caso por caso la pertinencia de las entidades reconocidas para la edición digital final. Esta revisión permitirá no solo depurar los errores residuales del modelo, sino también asegurar una vinculación más precisa con bases de autoridades. No obstante, estos resultados preliminares no invalidan la tarea de reconocimiento de entidades de LatinCy cuyo desempeño es notable, sino que simplemente remarcan la necesidad de que ciertas tareas automáticas tienen que ser auditadas y supervisadas siempre por especialistas. Desde un punto de vista práctico, los porcentajes de coincidencia entre las entidades lematizadas reconocidas y los nombres propios registrados en los *indices nominum* resultaron consistentemente buenos. En el caso de Catulo, las

entidades PERSON lematizadas coincidieron en un 69.8 %, las LOC en un 72.4 % y las NORP en un 61.0 %. Para Propercio, los valores fueron aún más elevados en LOC (76.4 %) y también robustos en PERSON (69.6 %) y NORP (66.2 %). En Tibulo, los porcentajes fueron igualmente satisfactorios: 66.2 % para PERSON, 70.1 % para LOC y 53.4 % para NORP (figura 3). En definitiva, si bien persiste la necesidad de revisión manual en casos ambiguos, la lematización con LatinCy mejora la alineación con repertorios onomásticos tradicionales y aumenta el rendimiento de las tareas automatizadas de procesamiento y edición digital.

Figura 3. Porcentaje de entidades encontradas en los *indices nominum* según la forma buscada: palabra flexionada o lema. Elaboración propia.



5.3 Identificadores persistentes y autoridades externas para enriquecer la edición

38 En esta sección se describe el pasaje desde las entidades reconocidas en los poemas (personas, lugares y grupos) hacia recursos externos que permiten normalizar nombres y preparar una edición TEI interoperable. Para ello se emplean tres fuentes de uso extendido en humanidades

digitales: VIAF, que reúne autoridades bibliográficas internacionales; Pleiades, orientada a lugares del mundo antiguo y a su información geográfica; y Wikidata, una base de conocimiento en forma de grafo con entradas para múltiples tipos de entidades. En lugar de enlazar manualmente cada mención, se realizan consultas automáticas mediante APIs para recuperar, para cada entidad, un conjunto de posibles coincidencias junto con sus identificadores. Estos identificadores persistentes podrán incorporarse posteriormente a la edición en XML TEI —por ejemplo, mediante `@ref`— tras una revisión editorial. Asimismo, se almacenan las respuestas completas de las consultas para auditar resultados, ajustar criterios y asegurar la trazabilidad del proceso.

5.3.1 Autoridades VIAF e identificadores persistentes

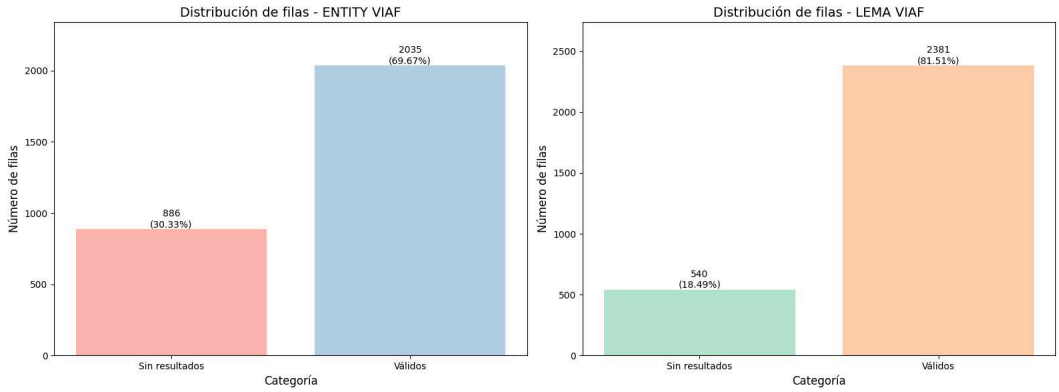
39 En la siguiente etapa se desarrolló un procedimiento de consulta automatizada utilizando la API proporcionada por VIAF.[17] El proceso partió del archivo CSV previamente generado con las entidades reconocidas por LatinCy, en el cual se incluían tanto la forma original como la forma lematizada de cada entidad, junto con su clasificación tipológica (persona, lugar o grupo). Para cada una de estas formas se realizaron búsquedas en la API de VIAF, con un límite de recuperación de hasta cinco coincidencias posibles. La información extraída —etiqueta y URI[18] de cada resultado — se almacenó en nuevas columnas del mismo archivo.

40 Además, como parte de una estrategia de trazabilidad y validación, se guardó el resultado completo de cada consulta en formato JSON. Esto no solo permite auditar los resultados, sino que también abre la posibilidad de realizar nuevos análisis con información más rica, como las relaciones de equivalencia con otras bases, descripciones ampliadas o variantes onomásticas.

41 Se procesaron un total de 2921 pares entidad/lema. En el caso de las entidades originales, el 69.67 % de las filas arrojaron al menos un resultado válido, mientras que un 30.33 % fueron identificadas explícitamente como *sin resultados* en VIAF. Por su parte, los lemas presentaron una tasa aún mayor de correspondencias: el 81.51 % de las entradas obtuvo al menos un resultado válido, frente al 18.49 % sin coincidencias registradas. Este comportamiento confirma que, en términos generales, la normalización lematizada favorece la vinculación con autoridades externas.

42 El análisis detallado de la distribución de resultados válidos muestra que la mayoría de las entidades y lemas con coincidencias lograron el máximo número de resultados posibles. Para las entidades originales, el 57.86 % de las filas con resultados obtuvo cinco coincidencias, mientras que en el caso de los lemas esa proporción se eleva al 69.94 % (figura 4).

Figura 4. Cantidad de resultados obtenidos en la consulta a la API de VIAF discriminados por término flexionado (*entity*) y lema. Elaboración propia.



5.3.2 Pleiades: lugares de la Antigüedad e identificadores persistentes

43 Para enriquecer la edición digital con información geográfica estructurada, se incorporaron datos provenientes de Pleiades, un proyecto comunitario (y grafo) de lugares antiguos que publica información autorizada sobre más de 36.000 ubicaciones históricas. Los datos fueron descargados desde su sitio oficial, que ofrece exportaciones periódicas en varios formatos, incluyendo JSON, CSV, KML y RDF. Para este proyecto, se utilizaron los archivos correspondientes a nombres (`names.csv`), puntos de localización (`location_points.csv`) y polígonos (`location_polygons.csv`), los cuales fueron integrados en un único conjunto de datos a través de una consulta que utilizaba un cruce por la clave `place_id`.

44 A partir del conjunto de datos base se extrajeron todos los términos únicos presentes en las columnas entity (correspondiente al término original flexionado) y lema. Estos términos fueron normalizados (convertidos a minúsculas y sin espacios innecesarios) y buscados en las columnas de las formas atestiguadas en textos (`attested_form`) y formas romanizadas, es decir, en idioma latín (`romanized_form`) de Pleiades. Para cada término se conservaron hasta cinco coincidencias como máximo. Las coincidencias encontradas se almacenaron en nuevas columnas del archivo CSV original, que incluyen tanto el URI de la entrada en Pleiades como sus geometrías asociadas, puntos o polígonos en un mapa (`geometry_point` y `geometry_polygon`).[19]

45 Para prever la posibilidad de que una entidad de lugar haya sido reconocida como persona, se buscaron todas las entidades que se obtuvieron con LatinCy (tanto personas como lugares). Por esta razón, el análisis de los resultados obtenidos a partir de la integración de los datos de Pleiades

sobre un total de 2921 filas arrojó un 13.63% de coincidencias válidas en las entidades (palabra flexionada) y un 18.56% en los lemas, un número relativamente bajo si se compara los resultados con los obtenidos de VIAF.

46 Un análisis por tipo de entidad muestra una mayor correspondencia con entidades geográficas (LOC), que arrojaron coincidencias en el 23.6% de los casos, en comparación con los grupos étnicos o gentilicios (NORP, 10.5%) y las personas (PERSON, 9.6%). La mayor cobertura para entidades geográficas se explica por la naturaleza de la base Pleiades. Esta información, almacenada directamente en el corpus en columnas diferenciadas por tipo y número de coincidencia, servirá como recurso para construir mapas interactivos y codificaciones geográficas en una futura edición digital TEI que no está incluida en este artículo.

5.3.3 Wikidata: búsqueda y recuperación de candidatos

47 La siguiente tarea consistió en el desarrollo de un flujo de trabajo que vinculara automáticamente las entidades obtenidas en los textos con entradas relevantes en Wikidata. Para ello, se reutilizó la lógica aplicada en fases anteriores con VIAF y Pleiades. El procedimiento partió del archivo que contenía las entidades identificadas con LatinCy (forma original y lematizada), junto con su tipo (PERSON, LOC, NORP). Para cada una de estas formas se realizó una consulta a la API de Wikidata utilizando el *endpoint* `wbsearchentities`,[20] que permite obtener posibles coincidencias basadas en el nombre. Las respuestas JSON fueron almacenadas localmente, y posteriormente procesadas para recuperar información clave: el identificador (QID), la URL, las fechas de nacimiento (P569) y defunción (P570), así como el tipo de entidad (P31, *instance of*).

48 Sobre cada par *entity/lemma* se ejecutaron consultas a la API de Wikidata mediante el *endpoint* `wbsearchentities`, recuperando hasta cinco candidatos por consulta. Para priorizar y ordenar los resultados se calculó un puntaje simple basado en tres señales: (i) coincidencia exacta entre el término buscado y la etiqueta del ítem, (ii) compatibilidad tipológica con la categoría esperada (PERSON/LOC/NORP) y (iii) coherencia cronológica cuando el ítem dispone de fechas, aplicando un umbral de ≤ 100 d. C. a partir de la fecha de defunción (P570) o, en su defecto, de nacimiento (P569). Con este esquema se definieron dos modalidades complementarias: una búsqueda estricta, que aplica filtros tipológicos y el umbral temporal cuando existen metadatos disponibles, y una búsqueda laxa, que omite estas restricciones para maximizar la cobertura, en particular en entidades sin fechas o sin tipificación clara (p. ej., deidades, personajes míticos). En ambos casos,

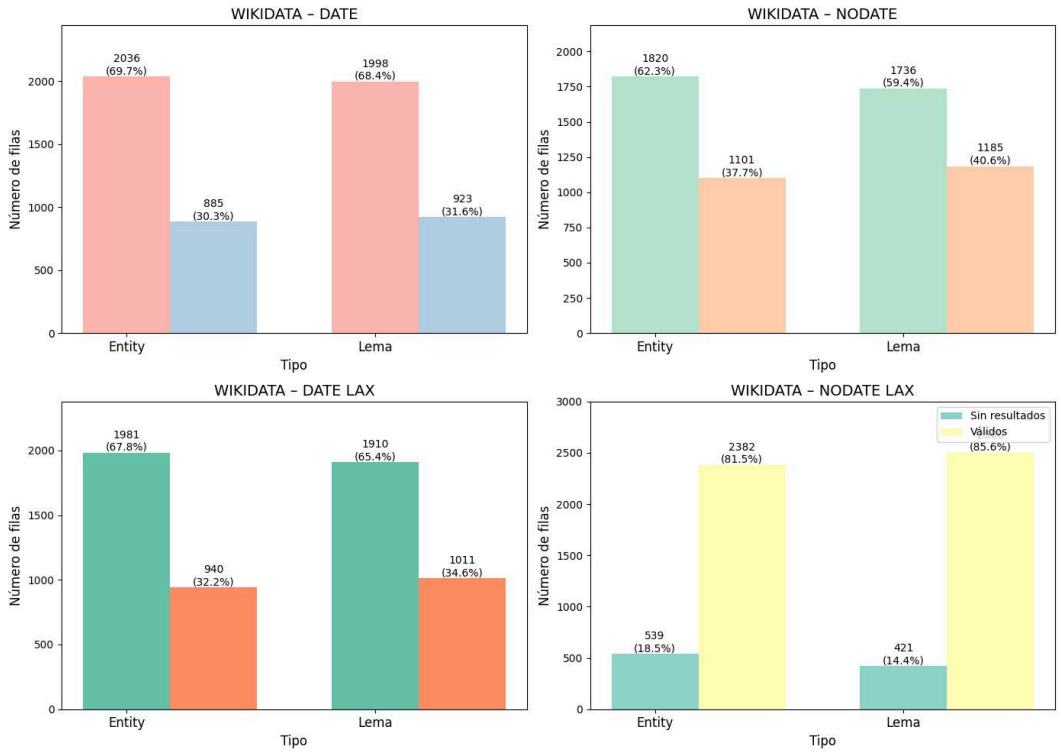
los candidatos se agrupan en “con fecha” y “sin fecha,” se ordenan por puntaje y se conservan los cinco mejor posicionados por grupo. Esta combinatoria (*entity*/*lemma* × estricta/laxa × con/sin fecha) constituye la base de la tabla 2.

Tabla 2. Resultados obtenidos en la consulta a la API de Wikidata discriminados por tipo de forma (*entity* o *lemma*) y por modalidad de búsqueda (estricta o laxa, con o sin metadato de fecha). Elaboración propia.

Consulta	Válidos	% Válidos	1 resultado	2 resultados
ENTITY - con fecha (estricta)	885	30.30%	884	1
ENTITY - sin fecha (estricta)	1101	37.69%	1083	18
ENTITY - con fecha (laxa)	940	32.18%	931	9
ENTITY - sin fecha (laxa)	2382	81.55%	1013	1369
LEMMA - con fecha (estricta)	923	31.60%	923	0
LEMMA - sin fecha (estricta)	1185	40.57%	1161	24
LEMMA - con fecha (laxa)	1011	34.61%	995	16
LEMMA - sin fecha (laxa)	2500	85.59%	958	1542

49 Se observa que la búsqueda estricta con fecha y tipo ofrece una precisión más alta, aunque limitada en cobertura. La búsqueda sin fecha aporta un incremento moderado, mientras que la variante laxa sin fecha es la que mayor cobertura proporciona, elevando los *matches* a más del 85 % en el caso de los lemas. También se verifica que en todas las variantes hay una fuerte concentración de coincidencias en un solo resultado por fila (más del 80 % en los casos estrictos), mientras que la variante `nodate_lax` permite acceder a múltiples resultados significativos: más del 34 % de las filas con 5 resultados válidos en *entity*, y 36.8 % en *lemma*. En conjunto, estos datos muestran que incorporar una segunda ronda de consultas más permisiva, sin restricciones cronológicas o tipológicas, resulta indispensable para mejorar la cobertura del sistema (figura 5). Aunque introduce ruido, permite reconocer entidades relevantes no tipificadas o sin información temporal, como deidades, héroes y nombres míticos frecuentes en la literatura grecolatina.

Figura 5. Cantidad de resultados obtenidos en la consulta a la API de Wikidata discriminados por término flexionado (*entity*) y lema (*lemma*), bajo distintos criterios de búsqueda: con filtro cronológico (estricta), sin filtro cronológico, búsqueda laxa con filtro cronológico y búsqueda laxa sin filtro. Elaboración propia.



50 Estos resultados no se interpretan como un enlazado definitivo, sino como un conjunto de candidatos priorizados que habilita una etapa posterior de curaduría editorial y su incorporación progresiva en TEI mediante identificadores persistentes. En una fase futura del proyecto se prevé ampliar la anotación de entidades nombradas añadiendo los identificadores persistentes procedentes de fuentes de autoridad y bases de conocimiento enlazadas. Siguiendo las directrices de TEI, esta anotación se implementaría empleando el atributo `@ref` para asociar a cada entidad uno o varios URI que remiten a registros de autoridad o conceptos bien definidos en grafos semánticos. TEI permite que el atributo `@ref` contenga una lista de punteros separados por espacios, lo que facilita la integración de múltiples referencias sin repetir el atributo, siempre que los prefijos utilizados estén documentados previamente en el `<teiHeader>` mediante `<listPrefixDef>`. Desde una perspectiva filológica, esta etapa implicará un proceso de curaduría posterior a la detección automática: las formas identificadas por el *pipeline* serán comparadas con los candidatos obtenidos mediante consultas a bases de conocimiento externas, ediciones anotadas y comentadas, y los resultados plausibles serán evaluados y validados por especialistas antes de su incorporación definitiva en la edición digital. Este enfoque resulta especialmente relevante en el caso de entidades literarias o poéticas que presentan ambigüedad semántica y múltiples correspondencias en los sistemas de autoridad. Un ejemplo representativo puede ser el nombre *Lesbia*, seudónimo literario utilizado por Cayo Valerio Catulo para referirse a su amante en el corpus poético. En Wikidata, Lesbia cuenta con un identificador específico como personaje literario asociado a la obra de Catulo (Q1235482). En una edición TEI enriquecida, una anotación futura podría consistir en la asociación de este identificador persistente al nombre propio mediante el atributo `@ref`, integrando así la entidad textual con un nodo semántico bien definido: `<persName ref="wd:Q1235482">Lesbia</persName>`.

5.3.4 Trazabilidad, organización de resultados y limitaciones

51 Con el fin de garantizar trazabilidad y reproducibilidad, los resultados de las consultas a Wikidata se almacenan en un archivo CSV que conserva explícitamente el origen de cada candidato según la estrategia de consulta (*entity*/*lemma*; estricta/laxa; con/sin fecha). Para cada candidato se preservan, además, el QID, la URL, los metadatos recuperados (incluyendo “instancia de” y fechas cuando están disponibles), la puntuación utilizada para ordenar resultados y la lista de alias, lo que permite auditar por qué un ítem fue priorizado y comparar el rendimiento relativo de

las estrategias. En paralelo, las respuestas JSON devueltas por `wbsearchentities` se almacenan íntegramente en disco para inspección posterior, depuración y posibles refinamientos del proceso de desambiguación.

52 Este diseño facilita una evaluación controlada de la fase de recuperación de candidatos —que en este trabajo se analiza comparando estrategias de consulta—, pero también impone limitaciones. En particular, la calidad del enlazado depende del reconocimiento y lematización previos: errores o inconsistencias en NER o en los lemas tienden a propagarse hacia las consultas y afectan tanto la cobertura como la precisión. Asimismo, las consultas laxas aumentan la recuperación de entidades relevantes sin metadatos completos, pero introducen inevitablemente más ruido, lo que refuerza la necesidad de una etapa posterior de curaduría y validación editorial por parte de los filólogos expertos.

5.4 Recuperación automatizada de Tópicos Amorosos en Catulo, Tibulo y Propercio utilizando el Diccionario de motivos amatorios en la literatura latina

53 Para la recuperación y marcado automático de tópicos amorosos se utilizó como recurso principal el *Diccionario de motivos amatorios en la literatura latina* (Moreno Soldevila 2011), una obra redactada en español que, sin embargo, incluye abundantes citas, ejemplos y fragmentos en latín. El objetivo del *pipeline* no fue procesar el contenido explicativo del diccionario, sino identificar, extraer y reutilizar únicamente los fragmentos en latín que correspondían a fragmentos presentes en los poemas del corpus.

54 En una etapa preliminar se evaluaron distintas estrategias computacionalmente más complejas para abordar este problema. Entre ellas se consideró la identificación automática de idioma con la finalidad de separar segmentos en español y latín, tanto a nivel de frase como de palabra, así como el uso de modelos de lenguaje para identificar y extraer automáticamente las porciones latinas relevantes. Sin embargo, las pruebas realizadas mostraron que estas aproximaciones introducían una complejidad elevada y resultados poco estables: la coexistencia de dos lenguas cercanas desde el punto de vista morfológico y léxico, junto con la presencia de latinismos plenamente integrados en el discurso académico en español, generaba confusiones frecuentes en las bibliotecas de identificación de idioma y en los modelos basados en aprendizaje profundo. En particular, se

observó que numerosos términos en español utilizados en el diccionario presentan una forma gráfica muy próxima o idéntica a sus equivalentes latinos, lo que dificulta una segmentación fiable y produce falsos positivos. Estas dificultades se acentuaban al trabajar con fragmentos breves, citas parciales o expresiones formularias, que son precisamente las unidades de mayor interés para el análisis temático. Además, el uso de modelos de lenguaje generativos, si bien prometedor, resultaba poco transparente desde el punto de vista metodológico y difícil de controlar en términos de reproducibilidad y trazabilidad de los resultados.

55 Inspirado en el trabajo de Lesk (1986),[21] uno de los primeros intentos computacionales de desambiguación semántica, se diseñó otra estrategia para minar textualmente el diccionario con una solución más simple y controlable, basada en la distancia de Levenshtein (1965), aplicada a la comparación de cadenas. Es decir, ya no nos interesaba identificar el idioma sino qué cadenas de caracteres de los poemas de nuestros autores latinos se encontraban en el diccionario. Esta estrategia permitió identificar fragmentos textuales de los poemas en el diccionario aun cuando existieran variaciones flexivas, ligeras divergencias ortográficas o adaptaciones propias de la redacción del diccionario, sin necesidad de realizar una segmentación lingüística o reconocimiento del idioma de cada fragmento.

56 El proceso consistió en la lectura y el procesamiento del diccionario: cada línea del texto del diccionario fue analizada para detectar, por medio de expresiones regulares, posibles encabezados temáticos escritos en mayúsculas (por ejemplo, “CELOS,” “DESPRECIO,” “MUERTE DEL AMADO”). A partir de esas líneas, se generaron n-gramas (secuencias de 3 a 15 palabras)[22] utilizando un procedimiento de tokenización y normalización.

57 Los poemas de los autores latinos se recorrieron línea por línea y cada línea se comparó con fragmentos del diccionario mediante dos estrategias complementarias: (a) comparación contra el fragmento completo y (b) comparación por n-gramas (subsecuencias de n palabras) para detectar coincidencias incluso cuando el diccionario conserva el pasaje incompleto o con separadores editoriales. Para medir cuán parecidos eran dos fragmentos, la distancia de Levenshtein (1965), una medida de “distancia de edición” que cuenta cuántos cambios mínimos (insertar, borrar o sustituir caracteres) hacen falta para transformar una cadena en otra, resultó sumamente apropiada. Este mismo principio se usa ampliamente en correctores ortográficos y en sugerencias de motores de

búsqueda del tipo “quizá quisiste decir...,” donde se intenta reconocer que dos secuencias son casi iguales aunque difieran por errores menores o por signos. Esto permitió tolerar diferencias pequeñas entre el poema y el diccionario, por ejemplo, los versos de Catulo:

> Vivamus, mea Lesbia, atque amemus,
> rumoresque senum severiorum
> (Catullus, 1893, Carm. 5.1–2)

58 Estos versos podrían aparecer en el diccionario como “Vivamus mea Lesbia atque amemus” (sin comas), como “Vivamus, mea Lesbia” (fragmento incompleto), como (“Vivamus, mea Lesbia)” (citado entre paréntesis) o incluso unidos con un separador editorial (/), por ejemplo, “Vivamus, mea Lesbia, atque amemus / rumoresque senum severiorum” que se suele utilizar para ahorrar espacio al citar textos en verso. En todos esos casos, el sistema sigue recuperando la coincidencia porque la distancia de edición entre variantes casi idénticas (con o sin comas, o con una barra `/`) es muy baja, y la búsqueda por n-gramas permite encontrar dentro de la línea la secuencia “vivamus mea lesbia atque amemus” aunque el diccionario la presente truncada o combinada con otro hemistiquio. Si la distancia entre el fragmento del poema y el del diccionario es menor o igual a un umbral configurable (por ejemplo, 1, 2 o 3 caracteres), se marca como coincidencia potencial para revisión editorial. Esta metodología permitió tolerar pequeñas diferencias en casos como en los que la notación particular del diccionario hiciera diferir levemente los textos.

59 Todas las coincidencias detectadas fueron almacenadas con información detallada: el fragmento buscado, el fragmento encontrado, la distancia de Levenshtein, el archivo y la línea del poema donde apareció, el tópico recuperado y el método de comparación utilizado (figuras 6 y 7). Posteriormente, se aplicó un filtro para conservar únicamente el fragmento más largo por línea y tópico, reduciendo los falsos positivos y simplificando la codificación.

60 En total, se extrajeron datos de 200 archivos TEI y se identificaron y anotaron un total de 371 tópicos para los poemas de Catulo, 450 para los de Tibulo y 730 para los de Propercio.

Figura 6. Vista de parte del archivo CSV en el que se almacenó la coincidencia de líneas entre el diccionario y los poemas utilizando varios umbrales de Levenshtein. Elaboración propia.

fragmento_buscado	fragmento_encontrado	distancia	archivo_texto	numero_linea_texto
vivamus mea lesbia atque	vivamus mea lesbia atque	0	Catulo_Carmen_005.txt	1
vivamus mea lesbia atque	vivamus mea lesbia atque	0	Catulo_Carmen_005.txt	1
vivamus mea lesbia atque amemus	vivamus mea lesbia atque amemus	0	Catulo_Carmen_005.txt	1
omnes unius aestimemus assis	omnes unius aestimemus assis	0	Catulo_Carmen_005.txt	3
omnes unius aestimemus assis	omnes unius aestimemus assis	0	Catulo_Carmen_005.txt	3
mi basia mille deinde centum	mi basia mille deinde centum	0	Catulo_Carmen_005.txt	7
da mi basia mille deinde centum	da mi basia mille deinde centum	0	Catulo_Carmen_005.txt	7
da mi basia mille deinde centum	da mi basia mille deinde centum	0	Catulo_Carmen_005.txt	7
mi basia mille deinde centum	mi basia mille deinde centum	0	Catulo_Carmen_005.txt	7

Figura 7. Vista de parte del archivo CSV en el que se almacenó la coincidencia de líneas entre el diccionario y los poemas y sus tópicos asociados. Elaboración propia.

linea_texto	seccion_diccionario	topico	metodo	umbral_usado
Vivamus, mea Lesbia, atque amemus,	DiccLine 1	INVITACIÓN AL DISFRUTE VITAL	ngrama	
Vivamus, mea Lesbia, atque amemus,	DiccLine 1	INVITACIÓN AL DISFRUTE VITAL	ngrama	1.0
Vivamus, mea Lesbia, atque amemus,	DiccLine 1	INVITACIÓN AL DISFRUTE VITAL	texto_completo	
omnes unius aestimemus assis.	DiccLine 1	INVITACIÓN AL DISFRUTE VITAL	texto_completo	
omnes unius aestimemus assis.	DiccLine 1	INVITACIÓN AL DISFRUTE VITAL	texto_completo	1.0
da mi basia mille, deinde centum,	DiccLine 3	BESOS	ngrama	1.0
da mi basia mille, deinde centum,	DiccLine 3	BESOS	texto_completo	
da mi basia mille, deinde centum,	DiccLine 5	AOJAMIENTO DE LOS AMANTES	texto_completo	
da mi basia mille, deinde centum,	DiccLine 5	AOJAMIENTO DE LOS AMANTES	ngrama	1.0

5.5 Diferentes opciones para el anotado de tópicos en el estándar XML-TEI

5.5.1 Codificación stand-off sobre archivos TEI

61 En una primera etapa del flujo de trabajo, se implementó un sistema de codificación temática en formato *stand-off* aplicado a archivos TEI. Para una primera etapa de codificación se filtraron aquellas coincidencias cuya distancia de Levenshtein no superaba un umbral de 1.

62 El *script* recorrió los archivos TEI correspondientes y, en cada uno, insertó un bloque `<standOff>` que contiene un `<spanGrp type="topicos">`. Dentro de este grupo, se incluye un elemento `<span>` por cada codificación temática identificada en el poema. Cada uno de estos elementos `<span>` especifica los atributos `@from` y `@to`, que indican el intervalo de líneas al que corresponde el tópico, y el atributo `@type`, que señala la categoría temática asignada (ejemplo 2). Además, en esta etapa

se agregan, en la declaración de la taxonomía TEI, que funciona como un vocabulario controlado de motivos, los identificadores estables (`@xml:id`) para cada motivo, de modo que las anotaciones puedan referenciarlos de manera unívoca (por ejemplo, con `@ana="#BESOS"`).

Ejemplo 2. Vista del archivo TEI resultante con codificación standoff. En la figura también puede observarse la entidad Lesbia reconocida automáticamente; aunque el modelo ha asignado erróneamente su tipo como lugar, lo que pone en evidencia la necesidad de una corrección y supervisión manual posterior. Elaboración propia.

```
<TEI>
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title xml:id="carmen005">Carmen 005</title>
    <author>Gaius Valerius Catullus</author>
    <respStmt>
     <name>Carlos Nusch</name>
     <resp>Edicion digital</resp>
    </respStmt>
    <respStmt>
     <name>Gabriel Calarco</name>
     <resp>Edicion digital</resp>
    </respStmt>
    <respStmt>
     <name>Gimena del Rio Riande</name>
     <resp>Edicion digital</resp>
    </respStmt>
   </titleStmt>
   <publicationStmt>
    <publisher>Proyecto Aetatis Amoris</publisher>
    <availability>
     <licence>
      <ref target="https://creativecommons.org/licenses/by/4.0/">CC BY 4.0</ref>
     </licence>
    </availability>
   </publicationStmt>
   <sourceDesc>
    <bibl>
```

```
    <author>Merrill, E. T.</author>
    <date>1893</date>
    <title>Catullus; edited by Elmer Truesdell Merrill</title>
    <publisher>Boston Ginn</publisher>
    <ptr target="http://archive.org/details/catulluseditedby00catuuoft"/>
   </bibl>
  </sourceDesc>
 </fileDesc>
 <encodingDesc>
  <appInfo>
   <application version="3.8.0" ident="latincy_la_core_web_lg">
    <label>la_core_web_lg v3.8.0</label>
    <ptr target="https://huggingface.co/latincy/la_core_web_lg"/>
   </application>
  </appInfo>
  <classDecl>
   <taxonomy xml:id="soldevila">
    <bibl>
     <author>Moreno Soldevila, R. (Ed.)</author>
     <date>2011</date>
     <title>Diccionario de motivos amatorios en la Literatura Latina</title>
     <publisher>Universidad de Huelva</publisher>
    </bibl>
    <category xml:id="AOJAMIENTO_DE_LOS_AMANTES">
     <catDesc>AOJAMIENTO DE LOS AMANTES</catDesc>
    </category>
    <category xml:id="BESOS">
     <catDesc>BESOS</catDesc>
    </category>
    <category xml:id="ENVIDIA_HACIA_LOS_AMANTES">
     <catDesc>ENVIDIA HACIA LOS AMANTES</catDesc>
    </category>
    <category xml:id="INVITACIÓN_AL_DISFRUTE_VITAL">
     <catDesc>INVITACIÓN AL DISFRUTE VITAL</catDesc>
    </category>
   </taxonomy>
  </classDecl>
 </encodingDesc>
```

```
   </teiHeader>
   <standOff>
    <spanGrp type="topics">
     <span xml:id="span-INVITACIÓN_AL_DISFRUTE_VITAL" from="1" to="6"
ana="#INVITACIÓN_AL_DISFRUTE_VITAL"/>
     <span xml:id="span-BESOS" from="7" to="13" ana="#BESOS"/>
     <span xml:id="span-AOJAMIENTO_DE_LOS_AMANTES" from="7" to="13"
ana="#AOJAMIENTO_DE_LOS_AMANTES"/>
     <span xml:id="span-ENVIDIA_HACIA_LOS_AMANTES" from="12" to="13"
ana="#ENVIDIA_HACIA_LOS_AMANTES"/>
    </spanGrp>
   </standOff>
   <text>
    <body>
     <lg type="poema">
      <l n="1">Vivamus, mea <placeName>Lesbia</placeName>, atque amemus,</l>
      <l n="2">rumoresque senum severiorum</l>
      <l n="3">omnes unius aestimemus assis.</l>
      <l n="4">soles occidere et redire possunt:</l>
      <l n="5">nobis, cum semel occidit brevis lux,</l>
      <l n="6">nox est perpetua una dormienda.</l>
      <l n="7">da mi basia mille, deinde centum,</l>
      <l n="8">dein mille altera, dein secunda centum,</l>
      <l n="9">deinde usque altera mille, deinde centum,</l>
      <l n="10">dein, cum milia multa fecerimus,</l>
      <l n="11">conturbabimus illa, ne sciamus,</l>
      <l n="12">aut ne quis malus invidere possit,</l>
      <l n="13">cum tantum sciat esse basiorum.</l>
     </lg>
    </body>
   </text>
  </TEI>
```

5.5.2 Codificación *flatten* o *in line* sobre archivos TEI

63 En una segunda modalidad de codificación, se optó por incorporar directamente la información temática en el cuerpo del texto mediante la inserción de atributos `@ana` dentro de los elementos `<l>` correspondientes, en lo que se denomina estrategia *flatten*. Esta modalidad tuvo como objetivo

ofrecer una codificación legible e inmediata dentro del propio flujo textual, útil tanto para editores humanos como para sistemas de procesamiento más simples que no implementen soporte para estructuras *stand-off*. En este caso se utilizó el atributo `@ana` correspondiente, con una codificación del tópico precedida por `#` (ejemplo 3).

Ejemplo 3. Vista del archivo TEI resultante con codificación flatten. Elaboración propia.

```
<TEI>
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title xml:id="carmen005">Carmen 005</title>
    <author>Gaius Valerius Catullus</author>
    <respStmt>
     <name>Carlos Nusch</name>
     <resp>Edicion digital</resp>
    </respStmt>
    <respStmt>
     <name>Gabriel Calarco</name>
     <resp>Edicion digital</resp>
    </respStmt>
    <respStmt>
     <name>Gimena del Rio Riande</name>
     <resp>Edicion digital</resp>
    </respStmt>
   </titleStmt>
   <publicationStmt>
    <publisher>Proyecto Aetatis Amoris</publisher>
    <availability>
     <licence>
      <ref target="https://creativecommons.org/licenses/by/4.0/">CC BY 4.0</ref>
     </licence>
    </availability>
   </publicationStmt>
   <sourceDesc>
    <bibl>
     <author>Merrill, E. T.</author>
     <date>1893</date>
```

```
    <title>Catullus; edited by Elmer Truesdell Merrill</title>
    <publisher>Boston Ginn</publisher>
    <ptr target="http://archive.org/details/catulluseditedby00catuuoft"/>
   </bibl>
  </sourceDesc>
 </fileDesc>
 <encodingDesc>
  <appInfo>
   <application version="3.8.0" ident="latincy_la_core_web_lg">
    <label>la_core_web_lg v3.8.0</label>
    <ptr target="https://huggingface.co/latincy/la_core_web_lg"/>
   </application>
  </appInfo>
  <classDecl>
   <taxonomy xml:id="soldevila">
    <bibl>
     <author>Moreno Soldevila, R. (Ed.)</author>
     <date>2011</date>
     <title>Diccionario de motivos amatorios en la Literatura Latina</title>
     <publisher>Universidad de Huelva</publisher>
    </bibl>
    <category xml:id="AOJAMIENTO_DE_LOS_AMANTES">
     <catDesc>AOJAMIENTO DE LOS AMANTES</catDesc>
    </category>
    <category xml:id="BESOS">
     <catDesc>BESOS</catDesc>
    </category>
    <category xml:id="ENVIDIA_HACIA_LOS_AMANTES">
     <catDesc>ENVIDIA HACIA LOS AMANTES</catDesc>
    </category>
    <category xml:id="INVITACIÓN_AL_DISFRUTE_VITAL">
     <catDesc>INVITACIÓN AL DISFRUTE VITAL</catDesc>
    </category>
   </taxonomy>
  </classDecl>
 </encodingDesc>
</teiHeader>
<text>
```

```
    <body>
     <lg type="poema">
      <l n="1" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">Vivamus, mea
<placeName>Lesbia</placeName>, atque amemus,</l>
      <l n="2" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">rumoresque senum severiorum</l>
      <l n="3" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">omnes unius aestimemus assis.</
l>
      <l n="4" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">soles occidere et redire
possunt:</l>
      <l n="5" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">nobis, cum semel occidit brevis
lux,</l>
      <l n="6" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">nox est perpetua una
dormienda.</l>
      <l n="7" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">da mi basia mille, deinde
centum,</l>
      <l n="8" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">dein mille altera, dein
secunda centum,</l>
      <l n="9" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">deinde usque altera mille,
deinde centum,</l>
      <l n="10" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">dein, cum milia multa
fecerimus,</l>
      <l n="11" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">conturbabimus illa, ne
sciamus,</l>
      <l n="12" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES
#ENVIDIA_HACIA_LOS_AMANTES">aut ne quis malus invidere possit,</l>
      <l n="13" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES
#ENVIDIA_HACIA_LOS_AMANTES">cum tantum sciat esse basiorum.</l>
     </lg>
    </body>
   </text>
  </TEI>
```

5.5.3 Codificación combinada: *flatten* + *stand-off*

64 Finalmente, se desarrolló una tercera modalidad de codificación que combinó simultáneamente los enfoques *flatten* y *stand-off* sobre los archivos TEI. Este enfoque híbrido da la posibilidad de facilitar tanto el procesamiento automático posterior como la inspección visual directa de

las codificaciones. El procedimiento consiste en insertar los atributos `@ana` directamente en los elementos `<l>` afectados por los tópicos (siguiendo la lógica *flatten*), al tiempo que se genera un bloque `<standOff>` con elementos `<spanGrp>` y `<span>` que codificaron la misma información temática de manera separada (*stand-off*) (ejemplo 4).

Ejemplo 4. Vista del archivo TEI resultante con codificación standoff y flatten. Elaboración propia.

```
<TEI>
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title xml:id="carmen005">Carmen 005</title>
    <author>Gaius Valerius Catullus</author>
    <respStmt>
     <name>Carlos Nusch</name>
     <resp>Edicion digital</resp>
    </respStmt>
    <respStmt>
     <name>Gabriel Calarco</name>
     <resp>Edicion digital</resp>
    </respStmt>
    <respStmt>
     <name>Gimena del Rio Riande</name>
     <resp>Edicion digital</resp>
    </respStmt>
   </titleStmt>
   <publicationStmt>
    <publisher>Proyecto Aetatis Amoris</publisher>
    <availability>
     <licence>
      <ref target="https://creativecommons.org/licenses/by/4.0/">CC BY 4.0</ref>
     </licence>
    </availability>
   </publicationStmt>
   <sourceDesc>
    <bibl>
     <author>Merrill, E. T.</author>
     <date>1893</date>
```

```
    <title>Catullus; edited by Elmer Truesdell Merrill</title>
    <publisher>Boston Ginn</publisher>
    <ptr target="http://archive.org/details/catulluseditedby00catuuoft"/>
   </bibl>
  </sourceDesc>
 </fileDesc>
 <encodingDesc>
  <appInfo>
   <application version="3.8.0" ident="latincy_la_core_web_lg">
    <label>la_core_web_lg v3.8.0</label>
    <ptr target="https://huggingface.co/latincy/la_core_web_lg"/>
   </application>
  </appInfo>
  <classDecl>
   <taxonomy xml:id="soldevila">
    <bibl>
     <author>Moreno Soldevila, R. (Ed.)</author>
     <date>2011</date>
     <title>Diccionario de motivos amatorios en la Literatura Latina</title>
     <publisher>Universidad de Huelva</publisher>
    </bibl>
    <category xml:id="AOJAMIENTO_DE_LOS_AMANTES">
     <catDesc>AOJAMIENTO DE LOS AMANTES</catDesc>
    </category>
    <category xml:id="BESOS">
     <catDesc>BESOS</catDesc>
    </category>
    <category xml:id="ENVIDIA_HACIA_LOS_AMANTES">
     <catDesc>ENVIDIA HACIA LOS AMANTES</catDesc>
    </category>
    <category xml:id="INVITACIÓN_AL_DISFRUTE_VITAL">
     <catDesc>INVITACIÓN AL DISFRUTE VITAL</catDesc>
    </category>
   </taxonomy>
  </classDecl>
 </encodingDesc>
</teiHeader>
<standOff>
```

```
    <spanGrp type="topics">
     <span xml:id="span-INVITACIÓN_AL_DISFRUTE_VITAL" from="1" to="6"
ana="#INVITACIÓN_AL_DISFRUTE_VITAL"/>
     <span xml:id="span-BESOS" from="7" to="13" ana="#BESOS"/>
     <span xml:id="span-AOJAMIENTO_DE_LOS_AMANTES" from="7" to="13"
ana="#AOJAMIENTO_DE_LOS_AMANTES"/>
     <span xml:id="span-ENVIDIA_HACIA_LOS_AMANTES" from="12" to="13"
ana="#ENVIDIA_HACIA_LOS_AMANTES"/>
    </spanGrp>
   </standOff>
   <text>
    <body>
     <l n="1" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">Vivamus, mea <placeName>Lesbia</
placeName>, atque amemus,</l>
     <l n="2" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">rumoresque senum severiorum</l>
     <l n="3" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">omnes unius aestimemus assis.</
l>
     <l n="4" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">soles occidere et redire
possunt:</l>
     <l n="5" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">nobis, cum semel occidit brevis
lux,</l>
     <l n="6" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">nox est perpetua una
dormienda.</l>
     <l n="7" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">da mi basia mille, deinde
centum,</l>
     <l n="8" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">dein mille altera, dein
secunda centum,</l>
     <l n="9" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">deinde usque altera mille,
deinde centum,</l>
     <l n="10" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">dein, cum milia multa
fecerimus,</l>
     <l n="11" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">conturbabimus illa, ne
sciamus,</l>
     <l n="12" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES
#ENVIDIA_HACIA_LOS_AMANTES">aut ne quis malus invidere possit,</l>
     <l n="13" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES
#ENVIDIA_HACIA_LOS_AMANTES">cum tantum sciat esse basiorum.</l>
    </body>
```

```
  </text>
</TEI>
```

5.6 Visualizaciones a partir de los datos extraídos y generados

65 Las tareas de procesamiento del lenguaje natural y codificación en XML TEI anteriores posibilitaron diferentes tipos de exploración aplicando métodos de lectura distante sobre el corpus. Por ejemplo, se pudo obtener una visión general sobre la distribución y frecuencia de los tópicos temáticos anotados en los poemas de Catulo, Tibulo y Propercio, gracias al desarrollo de un módulo de procesamiento que recorre los archivos TEI previamente anotados en formato *stand-off*.

66 El *script* extrajo, para cada archivo, el nombre del autor (a partir del nombre del archivo) y los tópicos registrados en los elementos `<span>` dentro del grupo `<spanGrp type="topicos">`. Cada aparición de un tópico se contabiliza según su atributo `@type`. Por medio de este procedimiento se obtuvo una visión cuantitativa de las ocurrencias por autor y por tipo de tópico (figuras 8, 9, y 10).

Figura 8. Vista de tópicos presentes en Catulo registrados en los elementos `<span>`. Elaboración propia.

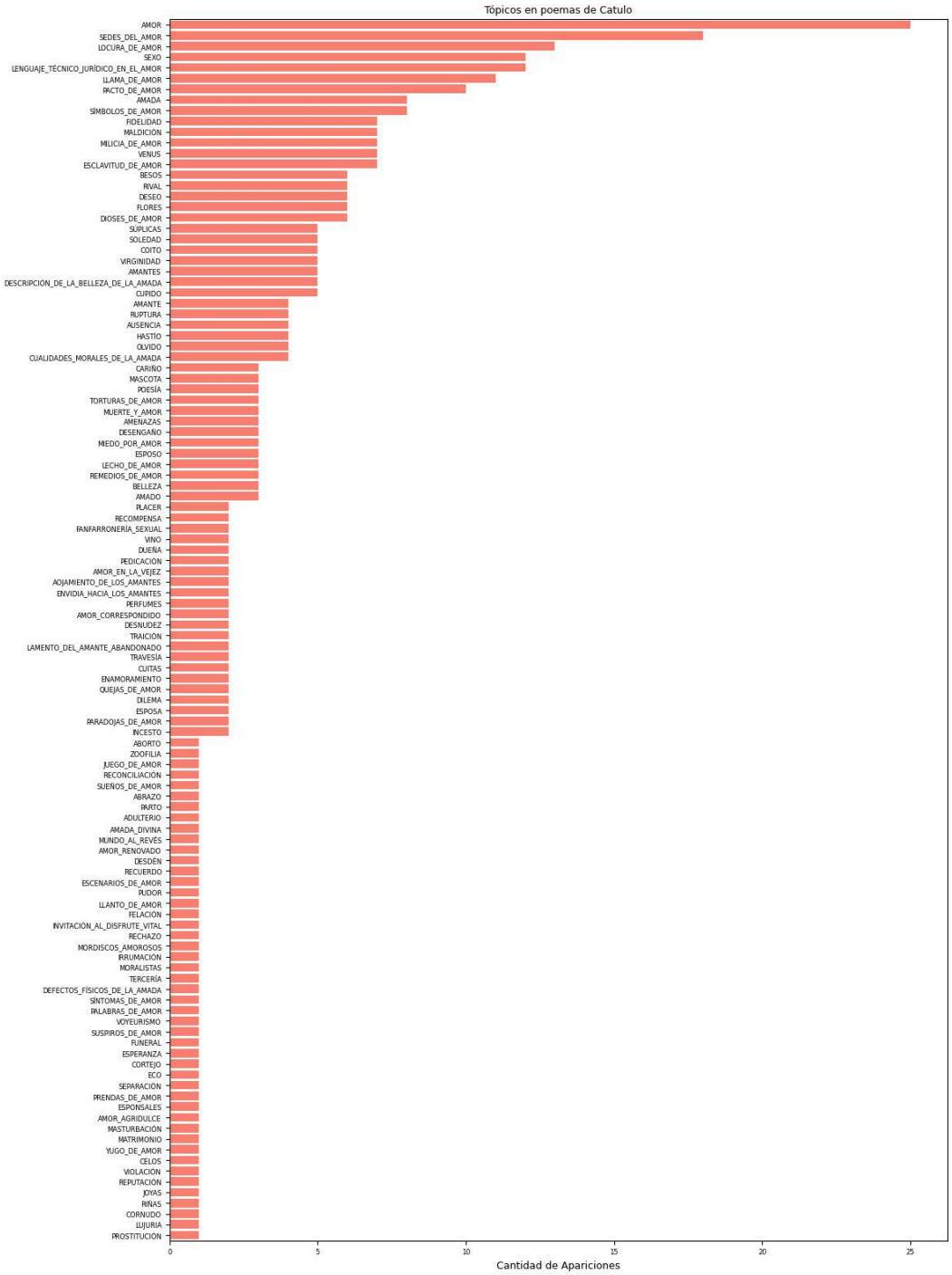


Figura 9. Vista de tópicos presentes en Tibulo registrados en los elementos `<span>`. Elaboración propia.

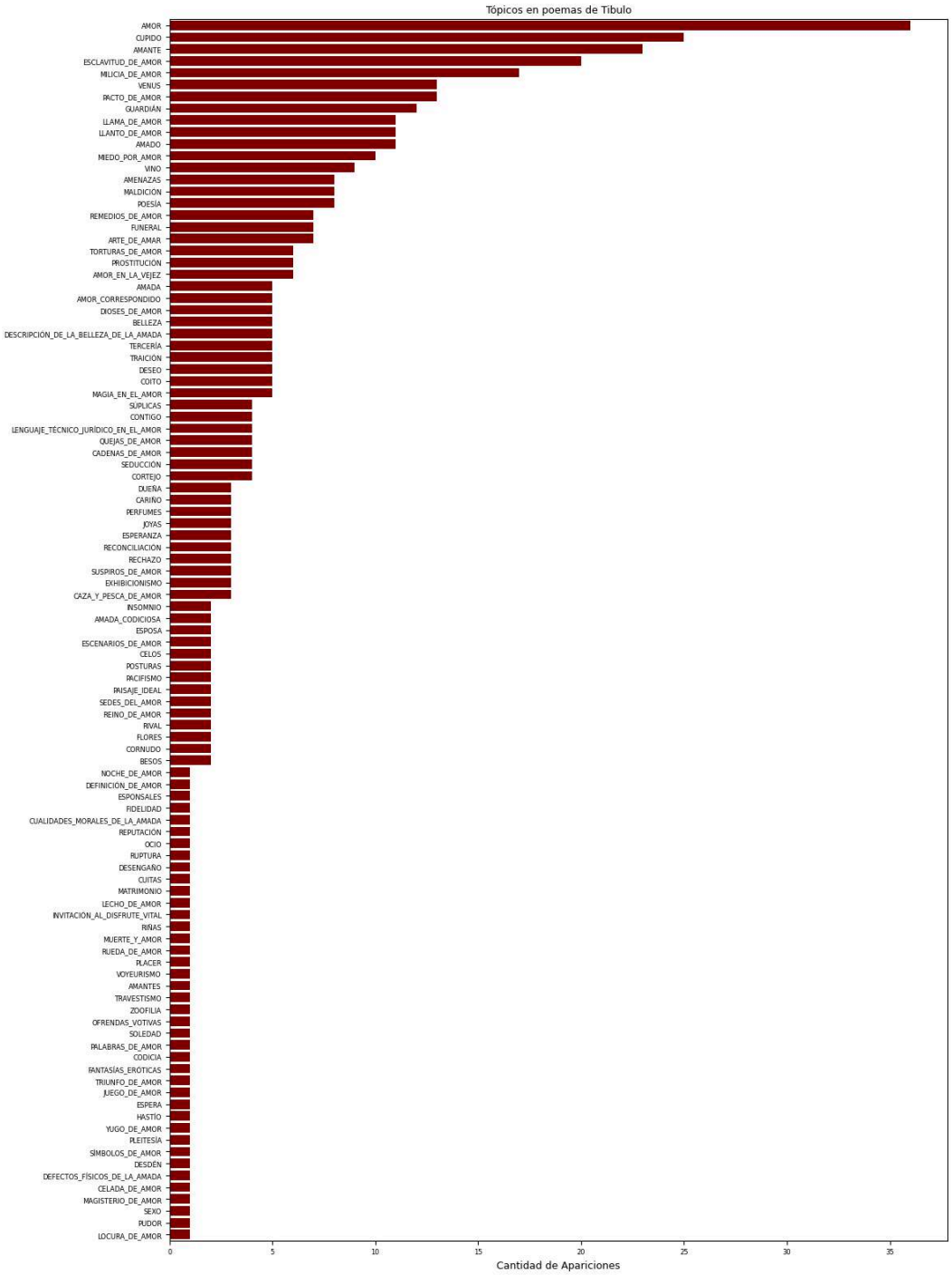
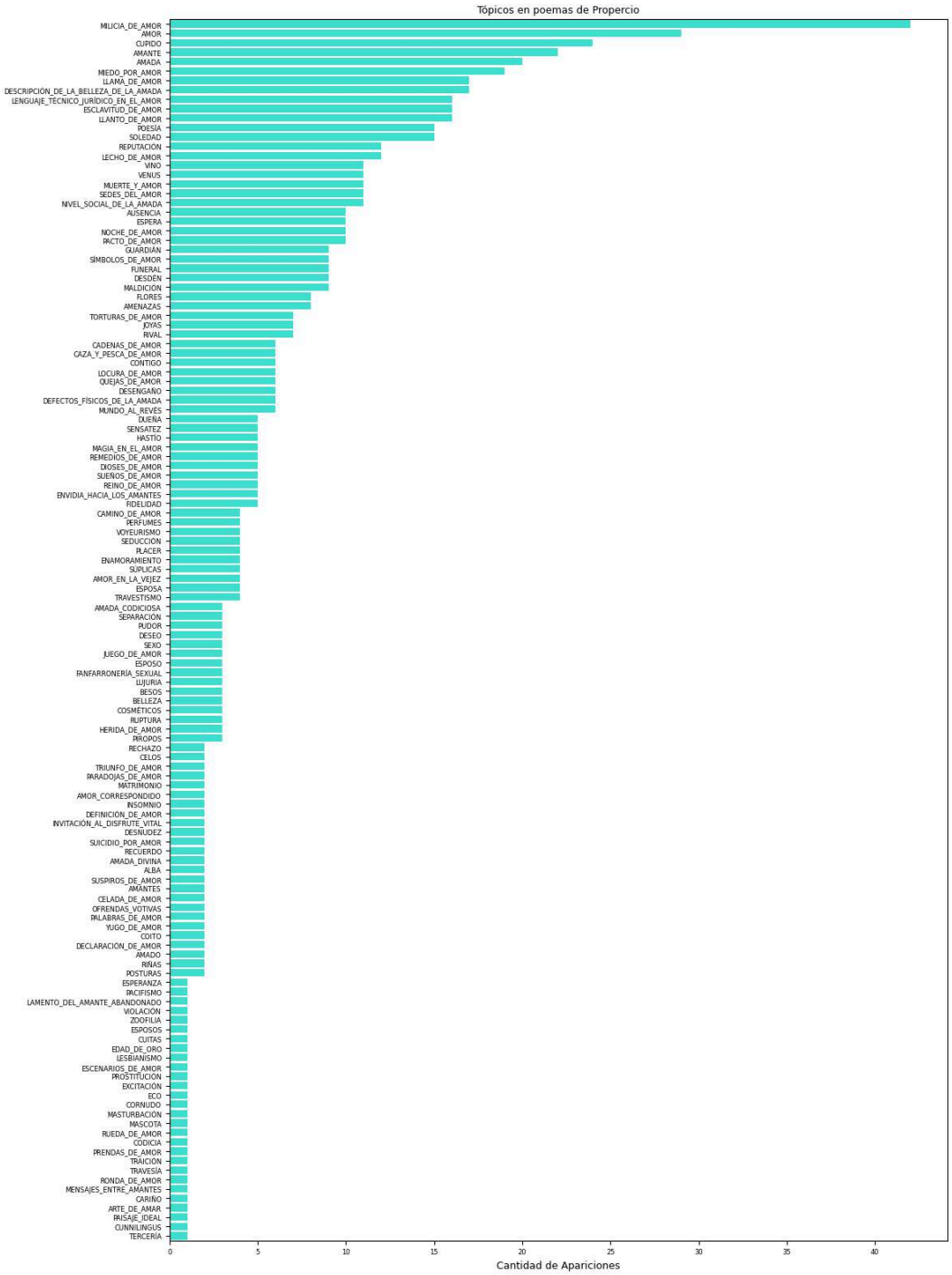


Figura 10. Vista de tópicos presentes en Propercio registrados en los elementos . Elaboración propia.

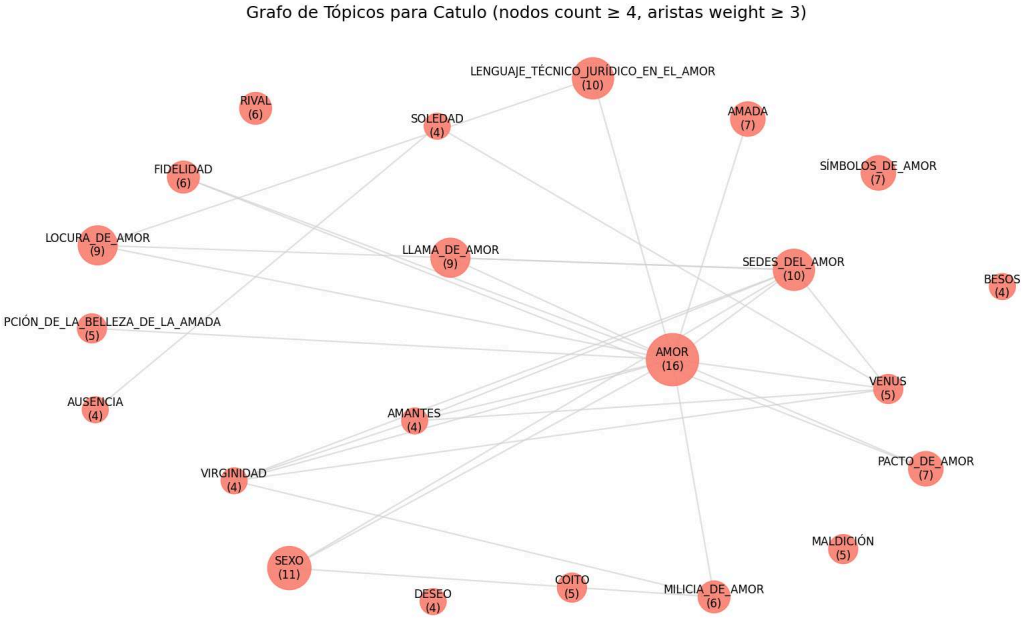


5.6.1 Visualización de coocurrencias temáticas mediante grafos

67 Otra metodología que permitió tener una idea general acerca de los tópicos encontrados fue la creación de grafos de coocurrencia. A partir de los archivos TEI con codificaciones en formato *stand-off*, se extrajo para cada poema el conjunto de tópicos presentes, y se construyó un grafo en el que cada nodo representa un tópico, y cada arista une dos tópicos que aparecen juntos en al menos un poema del mismo autor. Para generar estos grafos, se utilizó la biblioteca NetworkX, filtrando los nodos según un umbral mínimo de ocurrencia (`threshold`) y las aristas según la frecuencia de coocurrencia mínima (`edge_threshold`). Se establecieron configuraciones visuales específicas por autor, incluyendo color de nodo, tamaño del grafo, fuente y multiplicadores de tamaño de nodo, con el objetivo de producir visualizaciones comparables pero ajustadas a las particularidades de cada corpus.

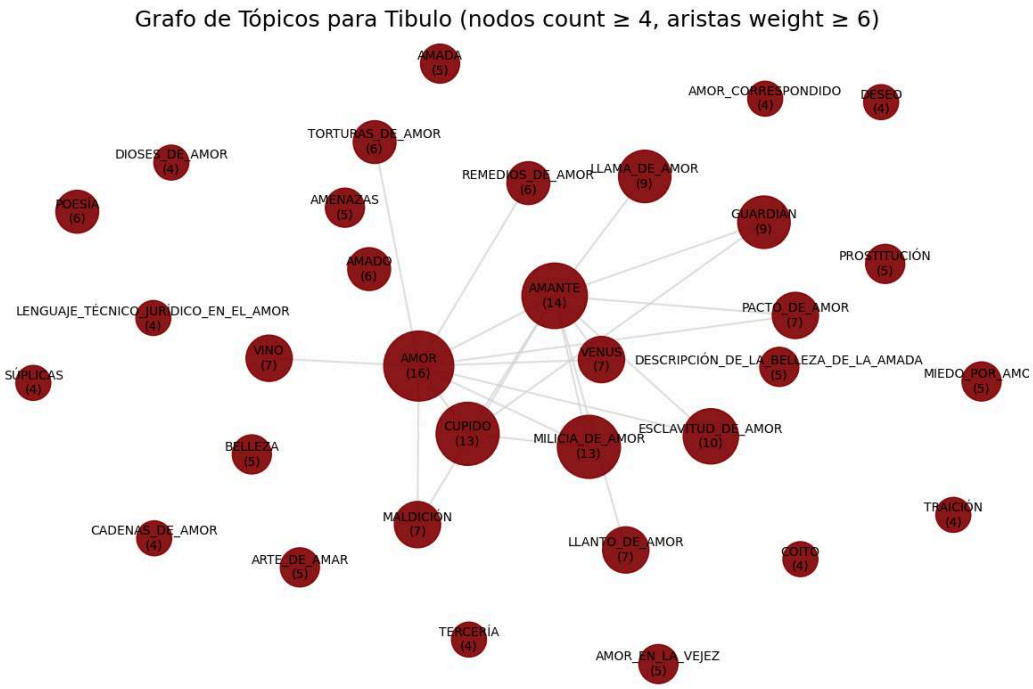
68 Los grafos resultantes permiten observar patrones de coocurrencia entre tópicos temáticos dentro del universo poético de cada poeta, revelando agrupamientos temáticos, tópicos dominantes y vínculos conceptuales frecuentes (figuras 11, 12, y 13).

Figura 11. Grafo de coocurrencia de tópicos en Catulo. Elaboración propia.



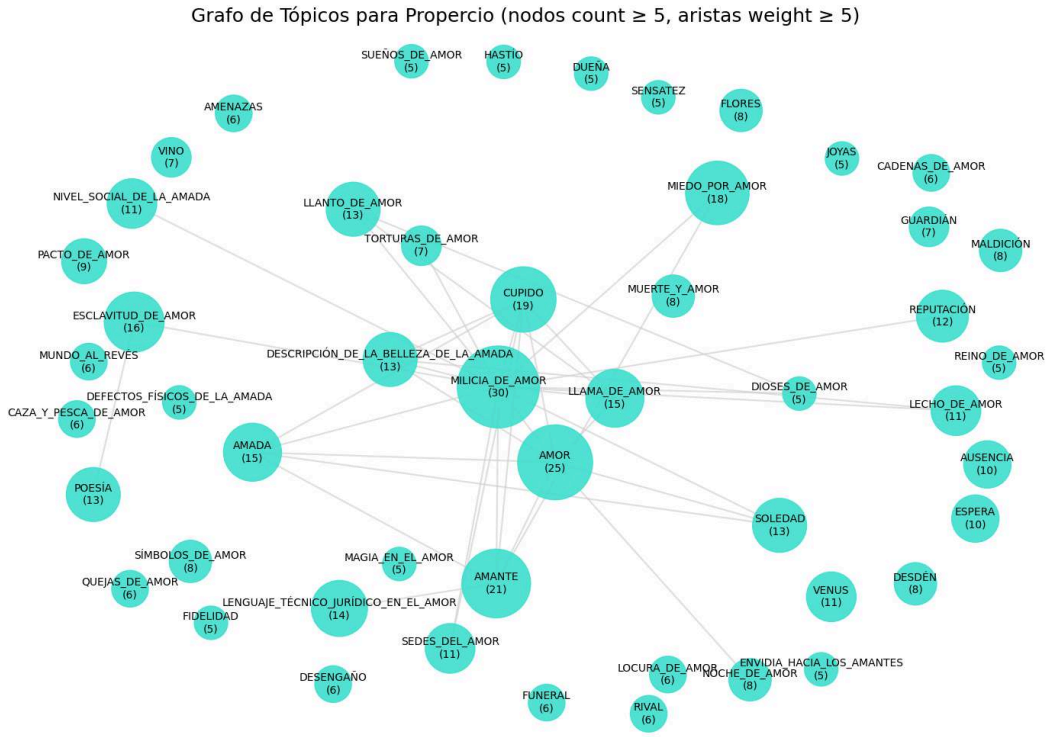
69 Los resultados muestran diferencias significativas tanto en la frecuencia total como en la diversidad temática entre los autores. Catulo registra un predominio de tópicos como *locura de amor* (13), *lenguaje técnico jurídico en el amor* (12), *sedes del amor* (esto es, las partes del cuerpo en las que el amor se expresa o habita) (18), *llama de amor* (11), y una notable presencia de *sexo* (12) y *maldición* (7). Con una centralidad marcada del *Amor* (concepto o personificación) que también se vincula bastante con *Venus*.

Figura 12. Grafo de coocurrencia de tópicos en Tibulo. Elaboración propia.



70 En Tibulo priman *milicia de amor* (17), *cupido* (25), *esclavitud de amor* (20), *venustas* y *vino*, junto con tópicos menos frecuentes en Catulo, como *funeral*, *magia en el amor* o *prostitución*. Hay una marcada cercanía entre *amante*, *amor*, *cupido*, conceptos típicamente elegíacos como la *milicia de amor* y la *esclavitud de amor*.

Figura 13. Grafo de coocurrencia de tópicos en Propercio. Elaboración propia.



71 Propercio es el poeta que presenta mayor densidad y variedad: *milicia de amor* (42), *llanto de amor* (16), *lenguaje técnico jurídico en el amor* (16), *soledad* (15), *poesía* (15) y *seducción* (4). Esta cuantificación no solo permite comparar el uso y peso relativo de ciertos motivos amorosos entre autores, sino también identificar elementos distintivos de cada poética. Por ejemplo, el alto número de tópicos en Propercio coincide con la apreciación de que su obra posee una mayor complejidad discursiva o una estructura más rica en motivos cuyo desarrollo fue posible no solo por su mayor tamaño, sino por las cualidades eruditas del poeta.

5.6.2 Grafo de coocurrencias de entidades nombradas

72 Se construyó también un grafo de coocurrencia de entidades para cada autor, donde los nodos representan entidades únicas (con su tipo y frecuencia), y las aristas reflejan la coocurrencia de dichas entidades dentro de un mismo texto.

73 Las visualizaciones generadas permiten observar con claridad los núcleos de entidades más frecuentemente mencionadas por cada autor, así como los vínculos semánticos entre ellas (figuras 14, 15, y 16).

74 Los grafos revelaron más patrones respecto del análisis cuantitativo anterior. En el caso de Catulo, se observa una gran centralidad de *Lesbia*, con múltiples conexiones a otras entidades, lo que confirma su papel dominante en el corpus. También aparecen numerosas variantes del propio nombre del poeta (*Catullus*, *Catullum*, *Catulli*, *Catullo*, *Catulle*) en consonancia con la omnipresente gravitación de su yo lírico y la constante construcción autorreferencial de su poesía. Diversas divinidades como *Iuppiter*, *Venus*, *Iovis* e *Iuno* y lugares míticos ponen en evidencia su imaginario mitológico que no solo se restringe a Venus como garante de justicia amatoria.

75 En el grafo de Tibulo, los núcleos más destacados son *Venus* y *Amor*, estrechamente conectados, lo que refuerza la idea de una tematización sistemática del amor. A diferencia de Catulo, hay menos duplicación de nombres y la red se presenta más equilibrada. *Messalla* destaca como figura histórica relevante, ya que era su mecenas y es nombrado repetidas veces en la obra. También aparece *Delia*, su amada.

76 El grafo de Propercio es el más denso y complejo. *Cynthia* aparece como nodo dominante, con una centralidad cuantitativa aún mayor que la de Lesbia en Catulo. La red presenta una abundante referencia a la mitología y la geografía (*Roma*, *Amor*, *Iuppiter*, *Venus*, *Troia*, *Romae*, *Graecia*), así como a figuras históricas como *Caesar*, *Pompeia* o el propio poeta. También se observan menciones a entidades colectivas como *Romanis*, *Remi*, *Lares* o *Musae*.

Figura 14. Grafo de coocurrencia de entidades utilizando palabras flexionadas en Catulo. Elaboración propia.

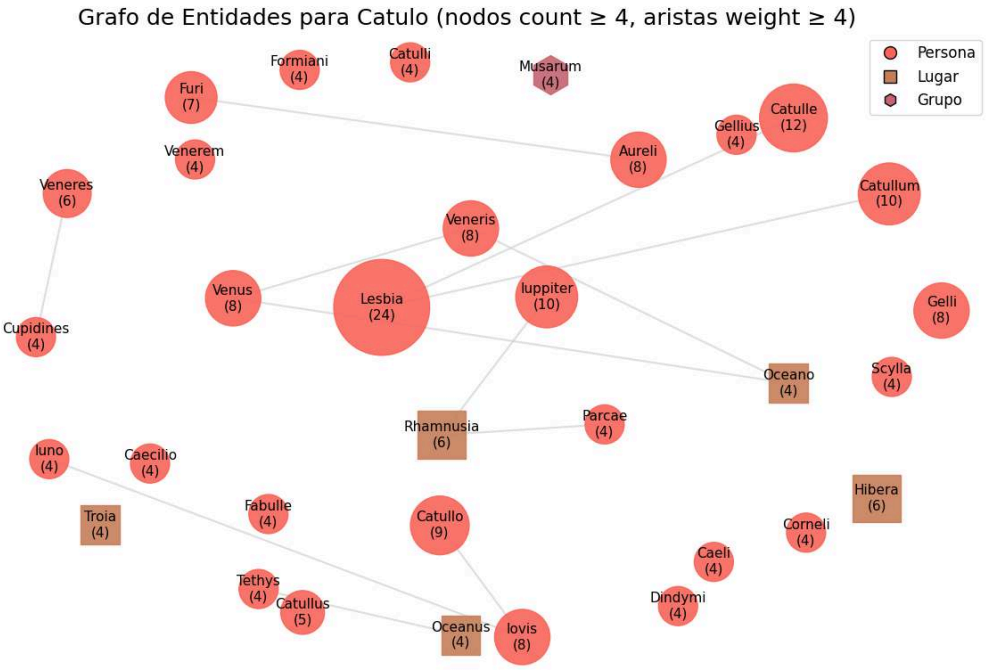


Figura 15. Grafo de coocurrencia de entidades utilizando palabras flexionadas en Tibulo. Elaboración propia.

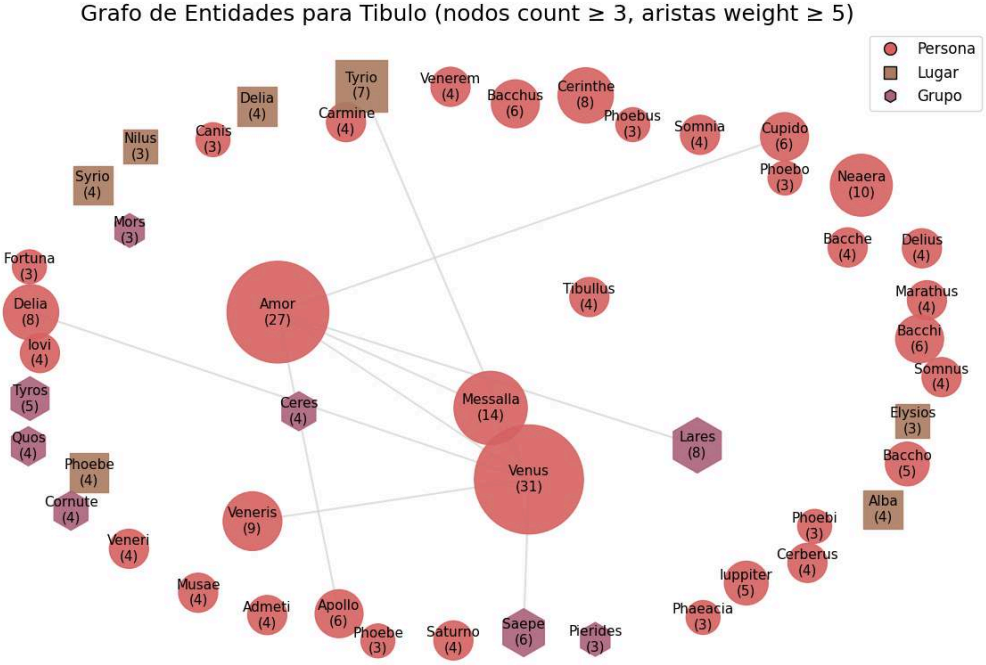
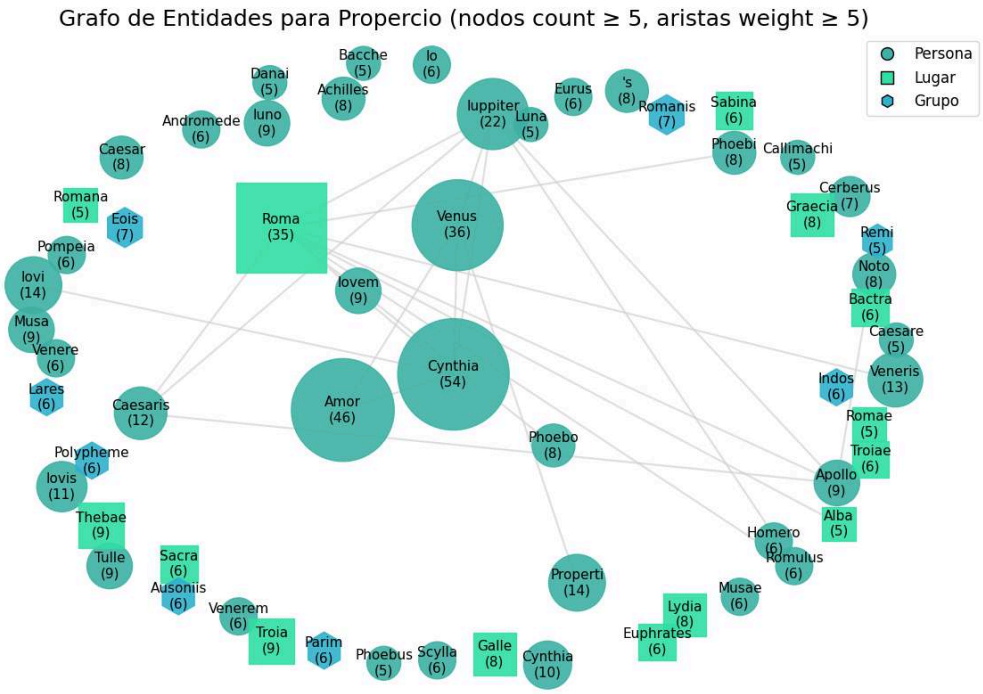


Figura 16. Grafo de coocurrencia de entidades utilizando palabras flexionadas en Propercio. Elaboración propia.



5.6.3 Grafo de coocurrencias de entidades nombradas lematizadas

77 Aprovechando la tarea de lematización realizada anteriormente, se elaboró otro grafo de coocurrencias por autor, donde los nodos representan entidades (lema y tipo), y las aristas representan la aparición conjunta de esas entidades en un mismo poema. Esta estrategia permitió unificar variantes morfológicas de una misma palabra, evitando repeticiones innecesarias de términos en los nodos del grafo. De este modo, entidades como *Iuppiter*, *Iovis* o *Iovi* se agrupan bajo un único lema.

78 Sin embargo, si una misma palabra fue identificada con distintos tipos de entidad —por ejemplo, si un lema aparece tanto como persona como lugar en distintos contextos—, se mantuvieron como nodos diferenciados. Esto asegura que las ambigüedades semánticas o funcionales de ciertos términos se reflejen adecuadamente en la estructura del grafo. Así, el análisis conserva tanto la compacidad léxica como la precisión tipológica de las entidades.

Figura 17. Grafo de coocurrencia de entidades utilizando palabras lematizadas en Catulo. Elaboración propia.

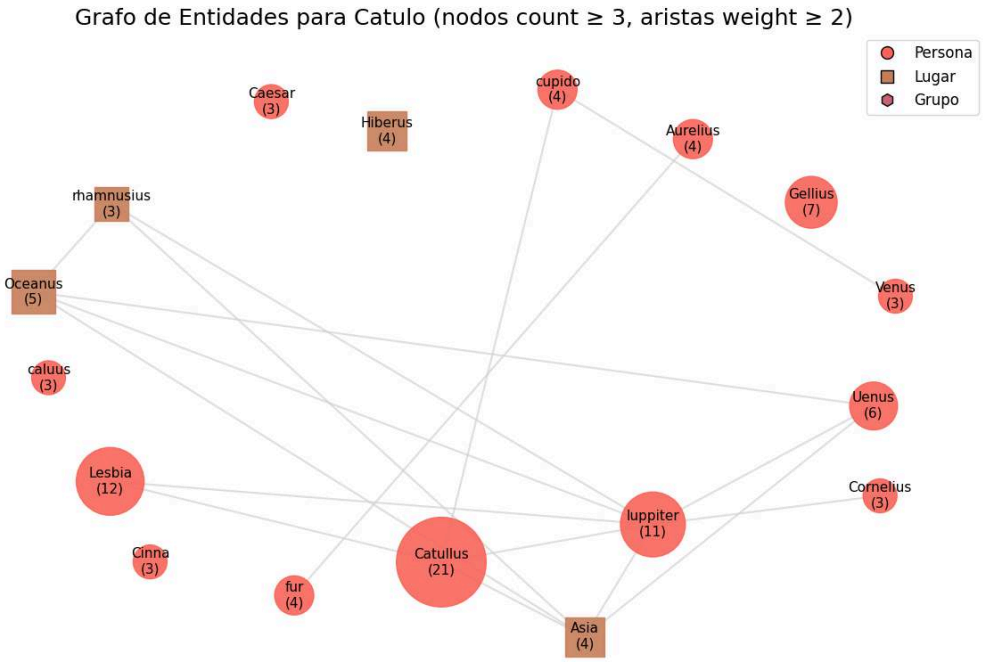


Figura 18. Grafo de coocurrencia de entidades utilizando palabras lematizadas en Tibulo. Elaboración propia.

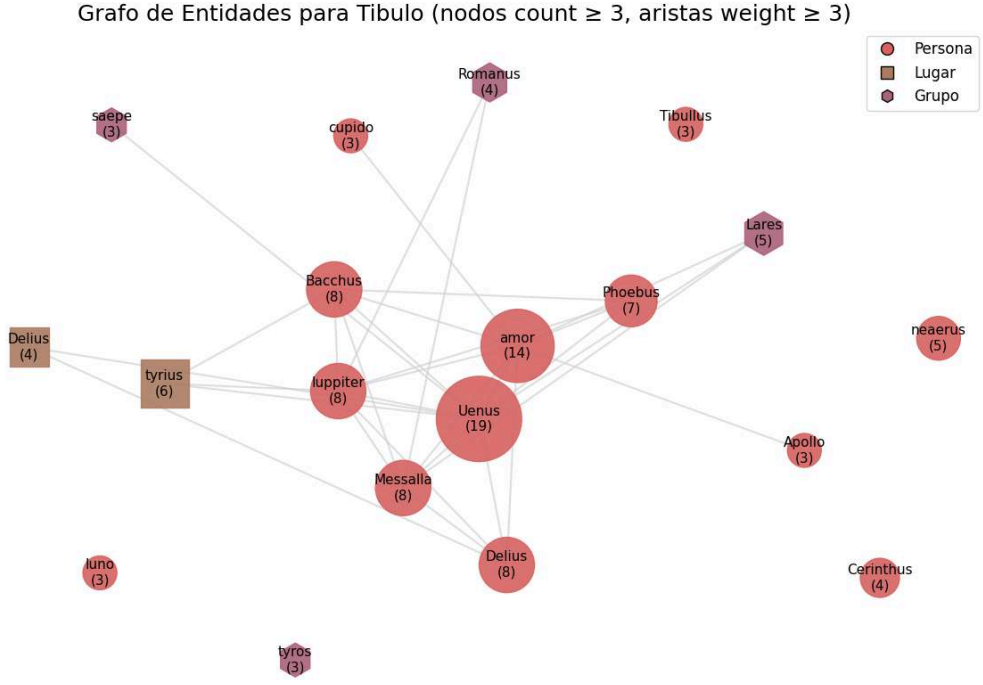
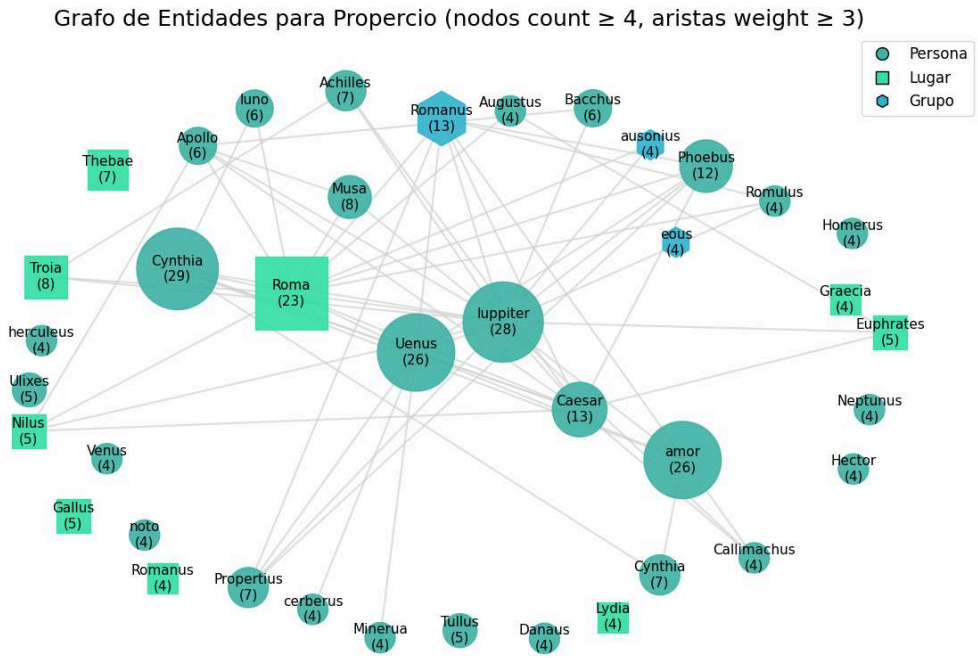


Figura 19. Grafo de coocurrencia de entidades utilizando palabras lematizadas en Propercio. Elaboración propia.



79 Los grafos generados a partir de los datos lematizados permiten observar con mayor claridad la centralidad y las relaciones de coocurrencia temática entre entidades en los poemas. Gracias a la lematización, se evita la dispersión de entidades por variantes morfológicas, lo que permite representar con mayor fidelidad los núcleos semánticos reales en cada autor.

80 En el caso de Catulo, se destacan los nodos de *Catullus*, *Iuppiter*, *Oceanus*, *Venus* y *Lesbia*, con conexiones múltiples entre ellos. La entidad *Catullus* tiene un papel central, rodeada por deidades y figuras personales recurrentes, lo que concuerda con la fuerte proyección autobiográfica o autorreferencial en su poesía.

81 En Tibulo, los nodos de mayor peso y conexiones son *Venus*, *Amor*, *Iuppiter*, *Bacchus* y *Messalla*. Se nota un enfoque temático en figuras mitológicas del amor y del vino. El grafo sugiere una mayor simetría entre núcleos amorosos y vínculos sociales o religiosos.

82 En Propercio, el grafo es más denso y amplio. Se destacan *Cynthia*, *Roma*, *Iuppiter*, *Venus*, *Caesar* y *Amor*, lo que evidencia una red más compleja de relaciones entre figuras amorosas, políticas y religiosas.

6. Discusión: lectura cercana y lectura distante como un círculo virtuoso

83 Si bien los procedimientos desarrollados a lo largo de este trabajo demostraron un notable grado de eficacia en la automatización de tareas editoriales y analíticas, es importante subrayar que ninguno de ellos está exento de limitaciones. El grado de automatización y fiabilidad varía según el tipo de tarea. La generación automática de encabezados y el conteo estructural de versos se mostraron altamente reproducibles, con muy bajo margen de error.

84 El reconocimiento de entidades nombradas —aunque basado en modelos robustos como LatinCy — puede arrojar resultados incorrectos o incompletos, especialmente en presencia de ambigüedad morfológica, sintáctica o referencial. Una futura tarea de ajuste fino (*fine tuning*)[23] del modelo LatinCy o la complementación con otro tipo de representaciones vectoriales de mayor dimensión como las que proporciona Latin-BERT (Bamman y Burns 2020) podría arrojar mejores resultados. A esto se suman las dificultades propias de las consultas a bases de datos externas como VIAF, Pleiades y Wikidata, donde no solo pueden producirse errores o respuestas vacías, sino también una sobreabundancia de resultados que desplaza el problema desde la búsqueda manual hacia la necesidad de filtrado de resultados o curaduría posterior. No obstante, este paso representa un avance metodológico significativo: permite sentar las bases para futuros análisis intertextuales, comparativos o semánticos y ofrece un marco escalable para la vinculación con datos externos.

85 El volumen de resultados devueltos por las consultas automáticas permite dimensionar el problema y justificar decisiones metodológicas: el archivo CSV en el que se almacenaron los resultados de los pares *entity/lemma* reúne del orden de decenas de miles de combinaciones entidad–candidato que deberían revisarse manualmente para construir un *gold standard* completo. Por ello, en esta fase el énfasis se puso en optimizar la obtención y organización de candidatos (cobertura, trazabilidad y control de ruido evidente), más que en completar la desambiguación y selección definitiva, que se plantea como trabajo futuro combinando reglas adicionales, *re-ranking* con metadatos, y eventualmente modelos de lenguaje, junto con validación manual sobre subconjuntos representativos.

86 Del mismo modo, el método empleado para la recuperación automática de tópicos temáticos a partir del diccionario de Soldevila et al., aunque falible y propenso a falsas identificaciones o a la omisión de variantes significativas, se revela como una herramienta sumamente útil para un primer acercamiento al corpus. Su utilidad no reside en su infalibilidad, sino en su capacidad para ofrecer una base de trabajo inicial amplia, estructurada y reutilizable, que puede ser posteriormente afinada mediante lectura cercana atenta.

7. Conclusiones

87 El objetivo central de este trabajo fue explorar estrategias de automatización que permitieran ahorrar tiempo y reducir errores en tareas de codificación mecánicas, rutinarias o propensas a inconsistencias, como la generación de encabezados `<teiHeader>`, el conteo y la codificación de versos, y el preetiquetado de entidades nombradas y tópicos amorosos. Para ello, se desarrolló un conjunto amplio de procedimientos automáticos aplicados al corpus poético, alineados con estándares de codificación digital XML-TEI. Estas tareas incluyeron el reconocimiento automático y etiquetado estructural de entidades nombradas mediante el modelo LatinCy, con su posterior normalización lematizada. Además, la información estructurada permitió consultar bases de datos de autoridades como VIAF, Pleiades y Wikidata, recuperar identificadores persistentes y enriquecer las entidades reconocidas con metadatos externos, con la idea de promover la interoperabilidad y el potencial analítico de la edición resultante. Como resultado, se logró producir un primer borrador de edición digital en formato XML-TEI bien estructurado, que integra un encabezado `<teiHeader>` completo, el marcado automático de los versos mediante elementos `<l>` numerados, y un pre-marcado de entidades y tópicos amorosos. Aunque esta información requiere posteriormente la supervisión y validación experta, el procedimiento permite ahorrar una cantidad significativa de tiempo y esfuerzo, y proporciona una base sistemática y organizada que facilita las tareas críticas de edición, análisis textual y exploración comparativa.

88 En cuanto al reconocimiento de entidades nombradas y lematización con LatinCy y su aplicación como insumos para la búsqueda de identificadores persistentes y datos que enriquecieran la edición en bases de datos de autoridades, se deben hacer varias consideraciones. VIAF devuelve en total 21466 URIs distribuidas en 10 columnas (forma flexionada y lema, con hasta cinco candidatos por cada una). De este modo, cada uno de los 2921 pares entidad–lema recibe al menos un

candidato, con un promedio superior a siete URIs por par. Pleiades devuelve 2015 URIs repartidas en 10 columnas, correspondientes a 644 de los 2921 pares entidad–lema (≈22 %), con alrededor de tres propuestas por par en los casos donde se registran coincidencias. Por su parte, Wikidata devuelve 14612 URIs en 40 columnas (ocho estrategias de consulta, con hasta cinco candidatos cada una), cubriendo 2228 pares entidad–lema (≈76 %); en los pares donde hay resultados, el promedio se sitúa en torno a 6.5 candidatos por entidad.

89 En conjunto, el archivo de resultados reúne del orden de 38000 combinaciones entidad–candidato (celdas no vacías correspondientes a URIs de VIAF, Pleiades y Wikidata). Este volumen evidencia dos cuestiones: por un lado, que la fase actual del pipeline funciona principalmente como un mecanismo de recuperación amplia de candidatos; por otro, que la desambiguación y la selección del identificador más adecuado requieren un procedimiento sistemático de evaluación y validación filológica. En este sentido, un paso previo imprescindible consiste en controlar la calidad de las entidades y lemas producidos automáticamente por LatinCy, ya que errores o inconsistencias en el reconocimiento de entidades nombradas (NER) o en la lematización se propagan hacia las consultas a las APIs y afectan tanto la cobertura como la precisión del enlazado.

90 Por ello, proponemos como trabajo futuro (i) la construcción de un *gold standard* sobre un subconjunto representativo del corpus que permita evaluar, en primer lugar, el desempeño de LatinCy en NER y lematización dentro de este dominio textual, y (ii) una vez estabilizada esa base, medir de manera controlada el desempeño del *entity linking* y comparar estrategias de filtrado y priorización. Sobre esa base, se evaluará la incorporación de métodos adicionales de reranking (por ejemplo, ponderando tipo de entidad, coincidencia nominal y compatibilidad cronológica) y de entity linking asistido por modelos de lenguaje (LLMs), utilizados como componentes de apoyo para reordenar candidatos y proponer enlaces plausibles a partir del contexto textual, manteniendo siempre la decisión final bajo supervisión editorial humana.

91 El procedimiento semiautomático de codificación temática de tópicos amorosos difiere sustancialmente del aplicado anteriormente para entidades nombradas, ya que implicó el desarrollo de un sistema de detección de patrones basado en n-gramas y distancia de Levenshtein, con umbrales controlados para minimizar falsos positivos. Los tópicos recuperados fueron anotados en los archivos TEI utilizando tres estrategias: codificación externa (*stand-off*), codificación directa en línea (*flatten*) y una modalidad combinada, que integra ambas. Esta tarea

permitió ahorrar tanto el tiempo de codificación como el de revisión exhaustiva del diccionario, aunque de ningún modo reemplaza esta última. Resta la tarea de revisar poema por poema y tópico por tópico recuperado para corroborar que las coincidencias hayan sido correctas y tengan sentido. La importancia de este método simple radica en que los tópicos codificados automáticamente se apoyan en la clasificación propuesta por un diccionario de gran prestigio académico y científico y no se utilizan modelos de lenguaje propensos al error o a la alucinación.

92 Por medio de esta tarea, no obstante, se registraron y procesaron datos de 200 archivos TEI, con una notable densidad de codificaciones de tópicos (371 en Catulo, 450 en Tibulo y 730 en Propercio), lo que permitió avanzar hacia una edición digital enriquecida y semánticamente estructurada, que a su vez habilita nuevas formas de visualización, análisis y exploración comparativa del discurso amoroso. Los datos obtenidos también serán reutilizados en tareas de clasificación supervisada, análisis estilométrico o generación de redes intertextuales con bases externas como VIAF o Wikidata.

93 El análisis conjunto de tópicos temáticos y redes de entidades en los poemas de Catulo, Tibulo y Propercio permitió trazar un mapa comparativo del imaginario amoroso elegíaco, revelando tanto continuidades como especificidades estilísticas y discursivas propias de cada autor.

94 Por último, cabe destacar una vez más que la investigación en lenguas antiguas mediante aprendizaje automático puede alcanzar un potencial enorme cuando especialistas en humanidades y especialistas en métodos computacionales colaboran estrechamente, combinando lectura distante y cercana. Este enfoque interdisciplinar permite abordar de forma más robusta los desafíos específicos de los textos antiguos al mejorar la interpretabilidad de los modelos y favorecer la replicabilidad de los resultados (Sommerschield et al. 2023). Por estas razones, ninguno de los métodos implementados pretende reemplazar la labor del especialista en filología, literatura o edición digital. Por el contrario, estos procedimientos buscan facilitar y enriquecer el trabajo de los investigadores, ofreciendo insumos organizados y comparables que permitan enfocar con mayor precisión la lectura cercana y la intervención crítica sobre los textos. La visualización de datos y los análisis cuantitativos que se desprenden de estos insumos tampoco deben ser considerados como un producto menor ya que pueden promover nuevas líneas interpretativas, favorecer el descubrimiento de patrones, y ofrecer una vía adicional para explorar el discurso amoroso elegíaco desde nuevas perspectivas.

BIBLIOGRAFÍA

Aguilar, Sergio Torres, Xavier Tannier, y Pierre Chastang. 2016. "Named Entity Recognition Applied on a Data Base of Medieval Latin Charters. The Case of Chartae Burgundiae." *3rd International Workshop on Computational History (HistoInformatics 2016)*. https://hal.science/hal-02407159/.

Bamman, David, y Patrick J. Burns. 2020. "Latin BERT: A Contextual Language Model for Classical Philology." arXiv:2009.10053. Preprint, arXiv, September 21. https://doi.org/10.48550/arXiv.2009.10053.

Bański, Piotr. 2010. "Why TEI Stand-off Annotation Doesn't Quite Work." Paper presented at Balisage: The Markup Conference. August 3. https://www.balisage.net/Proceedings/vol5/print/Banski01/BalisageVol5-Banski01.html.

Burns, Patrick J. 2023. "LatinCy: Synthetic Trained Pipelines for Latin NLP." arXiv:2305.04365. Preprint, arXiv, May 7. https://doi.org/10.48550/arXiv.2305.04365.

Cerrato, Lisa M., y Robert F. Chavez. s.f. "Perseus Classics Collection: An Overview." Accessed March 24, 2023. https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0053.

Ciotti, Fabio, Maurizio Lana, y Francesca Tomasi. 2014. "TEI, Ontologies, Linked Open Data: Geolat and Beyond." *Journal of the Text Encoding Initiative* 8 (December) https://doi.org/10.4000/jtei.1365.

Erdmann, Alexander, Christopher Brown, Brian Joseph, et al. 2016. "Challenges and Solutions for Latin Named Entity Recognition." In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, editado por Erhard Hinrichs, Marie Hinrichs, y Thorsten Trippel. The COLING 2016 Organizing Committee. https://aclanthology.org/W16-4012/.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Lesk, Michael. 1986. "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone." *Proceedings of the 5th Annual International Conference on Systems Documentation* (New York, NY, USA), SIGDOC '86, June 1, 24–26. https://doi.org/10.1145/318723.318728.

Levenshtein, V. 1965. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." *Soviet Physics. Doklady* 10 (8).

Lewis, C. S. 1936. *La alegoría del amor: un estudio sobre tradición medieval*. 2015th ed. Encuentro.

Lloyd, S. 1982. "Least Squares Quantization in PCM." *IEEE Transactions on Information Theory* 28 (2): 129–37. https://doi.org/10.1109/TIT.1982.1056489.

Merrill, Elmer Truesdell. 1893. *Catullus; Edited by Elmer Truesdell Merrill*. With Robarts - University of Toronto. Boston Ginn. http://archive.org/details/catulluseditedby00catuuoft.

Moreno Soldevila, Rosario. 2011. *Diccionario de motivos amatorios en la Literatura Latina*. Huelva. http://rabida.uhu.es/dspace/handle/10272/14398.

Moretti, Franco. 2013. *Distant Reading*. Verso.

Müller, Lucian. 1898. *Sex. Propertii Elegiae*. Teubner. https://archive.org/details/elegiaerecensuit00propuoft.

Nusch, Carlos Javier. 2021. "Las Edades del Amor: una propuesta para el proyecto Aetates Amoris destinado a la poesía amorosa." Tesis, Universidad Nacional de Educación a Distancia (UNED). https://doi.org/10.35537/10915/125629.

Nusch, Carlos Javier, Gimena Del Rio Riande, Leticia Cecilia Cagnina, Marcelo Luis Errecalde, y Leandro Antonelli. 2025. "Initial Explorations for Document Clustering Tasks in Latin Elegiac Poets." In *Collaboration in Knowledge Discovery and Decision Making*, editado por Vanessa Agredo-Delgado, Pablo H. Ruiz, y Carlos Augusto Meneses Escobar, vol. 2369. Communications in Computer and Information Science. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-91690-8_10.

Nusch, Carlos Javier, María Gimena del Rio Riande, Leticia Cagnina, Marcelo Luis Errecalde, y Rubén Leandro Antonelli. 2024. "Clustering Tasks and Decision Trees with Augustan Love Poets: Cohesion and Separation in Feature Importance Extraction." *CEUR Workshop Proceedings* 3834 (November). http://sedici.unlp.edu.ar/handle/10915/175050.

Postgate, John Percival. 1915. *Tibulli aliorumque carminum libri tres*. Scriptorum classicorum bibliotheca Oxoniensis. https://archive.org/details/tibullialiorumqu00tibuuoft.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1 (1): 81–106. https://doi.org/10.1007/BF00116251.

Ramsay, Stephen. 2011. *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press.

Sommerschield, Thea, Yannis Assael, John Pavlopoulos, et al. 2023. "Machine Learning for Ancient Languages: A Survey." *Computational Linguistics*, May 25, 1–44. https://doi.org/10.1162/coli_a_00481.

Viglianti, Raffaele. 2016. "Why TEI Stand-off Markup Authoring Needs Simplification." *Journal of the Text Encoding Initiative*, 10 (December). https://doi.org/10.4000/jtei.1838.

NOTAS

1 El código desarrollado para este trabajo se encuentra disponible en GitHub (https://github.com/KharolusIII/digital-edition-draft), incluye los *scripts* utilizados para la generación de archivos XML-TEI, ejemplos de entrada y salida, archivos CSV intermedios, documentación técnica y materiales

de apoyo para la reproducción del flujo de trabajo completo. Para facilitar la lectura, a lo largo del artículo se hará una definición somera de los conceptos más técnicos en notas al pie de página. Para obtener información más pormenorizada sobre las técnicas de Procesamiento del Lenguaje Natural y Minería de textos utilizadas, puede consultarse el repositorio y el Reporte Técnico incluido en el anexo, que también contiene un glosario de términos especializados..

2 El reconocimiento de entidades nombradas (NER, por sus siglas en inglés) es una tarea del procesamiento del lenguaje natural que consiste en identificar y clasificar automáticamente menciones de nombres propios en un texto, como personas, lugares, colectivos o instituciones.

3 VIAF es un proyecto cooperativo gestionado por OCLC (Online Computer Library Center), una organización sin fines de lucro dedicada a brindar servicios tecnológicos y de catalogación compartida a bibliotecas de todo el mundo. VIAF vincula los registros de autoridad de diversas bibliotecas nacionales e instituciones culturales, con el objetivo de consolidar las variantes de nombres personales, corporativos y geográficos bajo identificadores únicos y persistentes, para facilitar la normalización y la interoperabilidad en entornos bibliográficos y digitales. Véase: https://viaf.org/.

4 Pleiades es una comunidad y una base de datos abierta y colaborativa que proporciona identificadores únicos y metadatos detallados sobre lugares históricos del mundo grecorromano, con un enfoque especial en su localización geográfica, nombres antiguos y períodos de ocupación. Véase: https://pleiades.stoa.org/.

5 Wikidata es una base de conocimiento colaborativa, multilingüe y estructurada, mantenida por la Fundación Wikimedia, que permite representar entidades mediante propiedades y valores, y facilita su vinculación con otros proyectos y fuentes externas mediante identificadores persistentes. Véase: https://www.wikidata.org/.

6 La métrica utilizada en estos casos se denomina F-Score (o F1-Score) es una forma de evaluar qué tan bien funciona un sistema que identifica cosas automáticamente, como en este caso, nombres de personas o lugares en un texto. Esta medida combina dos aspectos importantes: 1) la precisión (*precision*) que indica cuántas de las cosas que el sistema encontró eran realmente correctas y 2) exhaustividad (*recall*) que indica cuántas de las cosas correctas que había en total, el sistema logró encontrar.

7 Los Campos Aleatorios Condicionales (CRF, por sus siglas en inglés) son modelos estadísticos utilizados para etiquetar secuencias de texto, ampliamente empleados en tareas como el reconocimiento de entidades nombradas.

8 spaCy es una biblioteca de código abierto para procesamiento del lenguaje natural que proporciona pipelines modulares para tareas como tokenización, análisis morfosintáctico, lematización y reconocimiento de entidades nombradas, ampliamente utilizada en entornos de investigación y producción. LatinCy es un conjunto de modelos y pipelines entrenados específicamente para el latín sobre el marco de spaCy, que integra árboles de dependencia universal y diversos corpus latinos para abordar las particularidades morfológicas y sintácticas de esta lengua. spaCy está disponible en https://spacy.io/ y LatinCy en https://spacy.io/universe/project/latincy.

9 Los árboles de dependencia universal (Universal Dependencies) son representaciones sintácticas que describen las relaciones gramaticales entre las palabras de una oración siguiendo un esquema estandarizado, lo que permite entrenar y comparar modelos lingüísticos entre distintas lenguas. El etiquetado morfosintáctico (*Part-of-Speech tagging*) es el proceso mediante el cual se asigna automáticamente a cada palabra de un texto una categoría gramatical, como sustantivo, verbo o adjetivo, junto con información morfológica relevante. La lematización es el proceso de reducir una palabra flexionada a su forma canónica o lema, lo que permite agrupar variantes morfológicas bajo una misma entrada léxica.

10 Anteriormente disponible en: http://www.geolat.it/. Lamentablemente el proyecto no parece seguir activo y el sitio ya no funciona.

11 En una etapa futura en la que se se realizará el marcado manual de los identificadores persistentes cosechados y filtrados, se agregará aquí también el detalle de las autoridades externas utilizadas.

12 Una API es una interfaz que permite acceder de forma automática y estructurada a los datos de un servicio externo. Aquí se emplean las APIs de VIAF, Wikidata y Pleiades para recuperar identificadores persistentes y metadatos de entidades reconocidas en los textos de forma masiva evitando el trabajo manual de buscar cientos de entidades, una por una, en sus respectivos sitios.

13 El modelo latincy_la_core_web_lg, versión 3.8.0 está disponible en: https://huggingface.co/latincy/la_core_web_lg.

14 Disponible en: https://www.perseus.tufts.edu/hopper/.

15 Disponible en: https://huggingface.co/latincy/la_core_web_lg.

16 Los *indices nominum*, o índices de onomásticos, son listados alfabéticos que recogen todos los nombres propios mencionados en una obra literaria, junto con sus ubicaciones en el texto. Se encuentran habitualmente en ediciones filológicas o libros académicos anotados.

17 Una API (*Application Programming Interface*) es una interfaz que permite la comunicación estructurada entre programas informáticos, y facilita, como en este caso, el acceso a los datos ofrecidos por VIAF de manera automática y en grandes cantidades. Se trata de una excelente opción para grandes consultas cuando se quiere evitar la consulta manual.

18 Una URI (sigla de *Uniform Resource Identifier*, en inglés) es una identificación única de un recurso en Internet. Sirve para nombrar, identificar o localizar recursos, como páginas web, imágenes, documentos, servicios o incluso conceptos.

19 Los términos entre paréntesis refieren a archivos y campos específicos del conjunto de datos de Pleiades usados para el cruce y la recuperación geográfica. `names.csv` reúne los nombres de los lugares (incluyendo variantes); `location_points.csv` contiene coordenadas puntuales (una ubicación expresada como punto) y `location_polygons.csv` contiene áreas o contornos (una ubicación expresada como polígono). Estos archivos se vinculan mediante `place_id`, que es el identificador interno del lugar en Pleiades. En los nombres, `attested_form` indica la forma atestiguada en las fuentes (tal como aparece registrada) y romanized_form una forma normalizada/romanizada para facilitar búsquedas (por ejemplo, en latín). En nuestros resultados, `geometry_point` y `geometry_polygon` almacenan la geometría asociada al lugar recuperado (punto o polígono). Finalmente, en nuestro DataFrame entity es la forma tal como aparece en el poema (a menudo flexionada) y lema es la forma canónica usada para agrupar variantes.

20 En una API, cada *endpoint* corresponde a una función o recurso concreto: buscar entidades, pedir detalles de un ítem, listar resultados, etc.; `wbsearchentities` es una "búsqueda interna" de Wikidata que permite, a partir de un nombre escrito (por ejemplo, un personaje o un lugar detectado en el texto), obtener automáticamente una lista de posibles coincidencias en Wikidata. Como resultado devuelve candidatos con su identificador, el nombre y una breve descripción, para que el editor pueda elegir luego cuál corresponde y, si lo desea, incorporar ese identificador persistente a la codificación TEI. Los elementos entre paréntesis indican los códigos internos con

los que Wikidata identifica propiedades específicas dentro de sus registros. Por ejemplo, P569 corresponde a la propiedad “fecha de nacimiento,” P570 a “fecha de defunción” y P31 a “es una instancia de” (es decir, el tipo de cosa que es la entidad, como `"persona"`, `"ciudad"`, `"dios"`, etc.). Estos códigos permiten recuperar esos datos de forma automática y consistente al procesar las respuestas en formato JSON.

21 Desde la premisa de que las palabras que aparecen juntas en un texto tienden a compartir un sentido coherente el algoritmo de Lesk compara las definiciones de diccionario (glosas) de una palabra ambigua con su contexto inmediato, y selecciona aquella acepción cuya definición comparte más palabras con el entorno. Aunque en este trabajo no se realizó una tarea de desambiguación de términos ni se utilizó dicho algoritmo, sí se desarrolló un sistema que recorría el diccionariopara utilizarlo como insumo en la recuperación de motivos amorosos en los poemas.

22 Un n-grama de palabra es una secuencia contigua de `n` palabras tomada de un texto. Por ejemplo, si `n` =2 (bigramas), “mea Lesbia” es un bigrama; si `n` =3 (trigramas), “mea Lesbia atque” es un trigrama. En el procesamiento del lenguaje natural, especialmente en enfoques clásicos previos al uso extendido de modelos neuronales, los n-gramas se utilizan como unidades básicas de representación y comparación.

23 *Fine tuning* (o “ajuste fino”) es una etapa de reentrenamiento en la que un modelo ya entrenado se adapta a una tarea o dominio específico mediante ejemplos anotados (un conjunto de referencia). En vez de empezar desde cero, se ajustan sus parámetros para que aprenda mejor los patrones del corpus objetivo (por ejemplo, cómo se mencionan personas o lugares en estos poemas latinos), lo que puede mejorar el reconocimiento de entidades, aunque requiere anotación manual previa y puede introducir sesgos si el conjunto de ajuste es pequeño o poco representativo.

AUTORES

CARLOS JAVIER NUSCH

Carlos Javier Nusch es Profesor y Licenciado en Letras por la Universidad Nacional de La Plata y Máster en Humanidades Digitales por la Universidad de Educación a Distancia de España. Ha publicado varios artículos sobre trabajo académico colaborativo, repositorios digitales, digitalización de patrimonio cultural, análisis del discurso político y literatura clásica, medieval y moderna. Trabaja en el Servicio de Difusión de la Creación

Intelectual (SEDICI) de la UNLP, en el Proyecto de Enlace de Bibliotecas (PREBI) y en el repositorio 38.14.CIC-Digital (CICPBA). Es Presidente de la Junta Directiva y coordina la Oficina de Relaciones Institucionales del Consorcio Iberoamericano para la Educación en Ciencia y Tecnología (ISTEC).

GABRIEL ALEJANDRO CALARCO

Gabriel Calarco es Licenciado en Letras por la Universidad de Buenos Aires (UBA) y se encuentra cursando el doctorado en Literatura en la misma universidad como becario del IIBICRIT-CONICET. Su proyecto de tesis está dedicado al estudio de la écfrasis en el Libro de Alexandre y al uso de herramientas de las Humanidades Digitales para la edición de textos medievales. También se desempeña como docente en la Diplomatura de Humanidades Digitales (UCES), como editor adjunto de la Revista de Publicaciones de la Asociación Argentina de Humanidades Digitales y como profesor de Literatura en la Educación secundaria.

GIMENA DEL RIO RIANDE

Gimena del Rio Riande es investigadora adjunta del Instituto de Investigaciones Bibliográficas y Crítica Textual (IIBICRIT) del Consejo Nacional de Investigaciones Científicas (CONICET), donde dirige el laboratorio de humanidades digitales (HD Lab). Es fundadora de la Asociación Argentina de Humanidades Digitales (AAHD) y directora de la Diplomatura en humanidades digitales de la Universidad de Ciencias Empresariales y Sociales (UCES). Desde 2018 es miembro de Board of Directors de la Text Encoding Initiative.

LETICIA CECILIA CAGNINA

Leticia Cecilia Cagnina es Doctora en Ciencias de la Computación, Magíster en Ciencias de la Computación y Licenciada en Ciencias de la Computación. Se desempeña como docente investigadora en la Universidad Nacional de San Luis (UNSL). Es Profesora Adjunta en el Departamento de Informática de la Facultad de Ciencias Físico-Matemáticas y Naturales de la UNSL. Además, es Investigadora Categoría Adjunto en la Carrera de Investigador Científico y Tecnológico del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Su experiencia profesional se enfoca en el campo de la Informática e Inteligencia Artificial, con especialidad en Procesamiento del Lenguaje Natural (PLN).

LEANDRO ANTONELLI

Leandro Antonelli es Licenciado en Informática por la Universidad Nacional de La Plata y se desempeña en el Laboratorio de Investigación e Informática Avanzada (LIFIA). También es Magíster en Ingeniería de Software y Doctor en Ciencias Informáticas por la misma universidad. Se ha desempeñado tanto en la academia como en la industria. En la academia ha atravesado distintas instancias de la docencia. Actualmente se desempeña como Jefe de Trabajos Prácticos en materias de grado y como profesor en materia de posgrado. También realizó investigación principalmente en ingeniería de requerimientos, con publicaciones en conferencias nacionales e internacionales, como así también en revistas.

MARCELO LUIS ERRECALDE

Marcelo Luis Errecalde es Profesor Exclusivo en la Universidad Nacional de San Luis (Argentina), y dirige el Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC) de la Facultad de Cs. Físico, Matemáticas y Naturales. Trabaja en Inteligencia Artificial, aprendizaje automático, minería de textos y Procesamiento del Lenguaje Natural. Colabora con diferentes grupos líderes de España, México, Alemania, Austria y Grecia en áreas como la calidad de la información en la web, detección de plagio, detección de depredadores sexuales en la web y determinación del perfil del autor (DPA). Actualmente, estudia la determinación del género, la edad, la orientación política y los rasgos de personalidad de los autores de documentos en la Web.