Revista de poética medieval, 33 (2019), pp. 51-76, ISSN: 1137-8905

POR UN ANÁLISIS DISTANTE Y PROFUNDO: UN CORPUS PILOTO DE LA POESÍA LÍRICA CASTELLANA DEL SIGLO DE ORO[*]

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

1. Por un análisis distante y profundo

Hace ya más de una década se planteó un nuevo modelo de análisis y estudio del texto literario (y del hecho literario en general) caracterizado por centrarse no tanto en el estudio en profundidad de uno o varios textos literarios concretos sino en el análisis general de gran cantidad de textos literarios más o menos representativos de un fenómeno literario común, un tema recurrente, un grupo de autores, una época o todo un periodo. Este nuevo enfoque se ha denominado *distant reading*[1] (en contraposición al *close reading*) o «macroanálisis»[2] (por comparación con el concepto de análisis «macro» de la macroeconomía). Sin entrar a discutir nomenclaturas o traducciones, en este artículo lo denominaremos simplemente «análisis distante».

Más que un modelo alternativo al análisis tradicional y profundo del texto literario, el enfoque que propone el análisis distante es perfectamente compatible con él[3]. El análisis distante tiene un objetivo panorámico: busca determinar los aspectos comunes, generales o recurrentes de todo un periodo o

[*] Artículo financiado por las ayudas Fundación BBVA a equipos de investigación científica, proyecto «Análisis distante de base computacional del soneto castellano del Siglo de Oro (ADSO)» (2016-2018). Enlace: <http://adso.gplsi.es> [consulta: 05/11/2019].

[1] Cfr. Franco Moretti, *La literatura vista desde lejos*, Barcelona, Marbot ediciones, 2007.

[2] Cfr. Matthew L. Jockers, *Macroanalysis.Digital Media and Literary History*, Illinois, University of Illinois Press, 2013.

[3] Cfr. Borja Navarro Colorado, «Complementariedad de los *Big Data* y los estudios literarios», en *#Nodos*, ed. de Gustavo Ariel Schwartz y Víctor Bermúdez, Pamplona, Next Door Publishers, pp. 460-464.

una época[4]. Más que por la especificidad literaria de una obra o un autor, el análisis distante se interesa, por lo común, por un amplio grupo de autores u obras. Es por esto por lo que ambos modelos son perfectamente compatibles: en la medida en que se pueda establecer lo común y general de un periodo se podrá determinar lo específico de una obra literaria con relación a ese periodo. Este puede ser tanto su contexto de producción (rasgos literarios generales de la época en que la obra fue creada) como su contexto de recepción (rasgos literarios generales de la época en que la obra es o fue leída)[5].

El análisis distante suele aplicar técnicas computacionales de análisis textual (*text mining*, procesamiento del lenguaje natural, *machine learning*, etc.) precisamente por la gran cantidad de texto literario que pretende analizar. Esta es una de sus principales características. De esta manera el análisis distante cubre un área macroanalítica inalcanzable para un análisis manual tradicional, una visión panorámica del hecho literario que sin la tecnología computacional no se podría realizar. La tecnología computacional actúa así como un microscopio o un telescopio para la ciencia natural, mostrando aspectos del objeto de análisis en niveles (en este caso «macros») inalcanzables para el ser humano.

Por esta misma razón, por la inmensa cantidad de texto literario a analizar, para el análisis distante se suelen utilizar, sobre todo, métodos de análisis computacional basados en modelos de aprendizaje «no supervisado».

Dentro del área del *text mining* y el aprendizaje automático se suelen establecer dos modelos generales: los modelos de aprendizaje supervisados y los modelos de aprendizaje no supervisados[6]. En el caso del procesamiento

[4] En cierta manera el modelo de análisis distante guarda bastante similitudes con los análisis tematológicos y comparatistas. Estos también analizan gran cantidad de obras diferentes de manera transversal para estudiar aspectos comunes a ellas (temas recurrentes, tópicos literarios, motivos, mitos, etc.). Hay, sin embargo, dos grandes diferencias tanto en la cantidad de obras analizadas como en los métodos utilizados. El análisis distante supone un aumento considerable en la cantidad de obras a analizar, pues analiza muchas más obras de las que una persona sola podría estudiar. Por esta razón el análisis de estas amplias colecciones de textos se realiza con modelos y técnicas computacionales. Sería interesante un estudio comparativo de ambos modelos, sobre todo para establecer ajustes terminológicos. Sobre este punto, véase Borja Navarro Colorado, «On Poetic Topic Modeling: Extracting Themes and Motifs from a Corpus of Spanish Poetry», en *Frontiers in Digital Humanities*, 5:15 (2018). DOI: 10.3389/fdigh.2018.00015.

[5] Sobre este punto, y tratando la poesía de Garcilaso, comenta Antonio García Berrio: «creo que el único camino útil y razonable para construir la periferia contextual de la cultura literaria de Garcilaso, o de cualquier otro de nuestros autores clásicos, hay que esperarlo a partir del tratamiento informático masivo sobre las canteras de cultura clásica y clasicista, para configurar con todo ello parámetros de interpretación global sobre el comportamiento individual o social en la cultura» (Antonio García Berrio, «Retórica figural. Esquemas argumentativos en los sonetos de Garcilaso», en *El centro en lo múltiple (selección de ensayos) II. El contenido de las formas (1985-2005)*, Barcelona, Anthropos, 2000, pp. 228-240).

[6] Cfr. Christopher D. Manning y Hinri Schütze, *Foundations of Statistical Natural Language Processing*, Massachusetts, MIT Press, 1999, pp. 232 y ss.

de texto, en ambos métodos el sistema aprende por sí mismo cómo analizar grandes cantidades de texto. La diferencia entre ellos radica en cómo hacen ese aprendizaje automático. Los métodos supervisados, para aprender cómo analizar un texto, parten de un corpus textual ya analizado por un experto. Estos métodos disponen de datos correctos para, a partir de ellos, analizar nuevos textos.

Por ejemplo, para realizar un sistema que clasifique automáticamente textos según su tema, un modelo de aprendizaje supervisado necesita una colección de textos (corpus) que hayan sido previamente clasificados a mano según su tema. A partir de esta clasificación humana, el algoritmo de análisis supervisado aprende de manera automática qué rasgos son propios de cada tema (por ejemplo, cuáles son los términos específicos de cada uno) y, con ello, clasifica nuevos textos.

Con los métodos no supervisados el sistema aprende cómo analizar automáticamente un texto prácticamente desde cero. Estos métodos parten de un corpus sin ningún tipo de anotación ni marca, solo texto: palabras separadas por espacios en blanco. Extraen información del texto mediante cálculos de frecuencia y procesos de inferencia. Así funciona, por ejemplo, el algoritmo LDA *Topic Modeling*[7], uno de los más utilizados para el análisis distante de amplios corpus de texto literario[8]. Este algoritmo extrae automáticamente temas o *topics* recurrentes de una amplia colección de textos a partir de las palabras que tienden a aparecer en los mismos textos o contextos. La idea intuitiva en la que se basa este y otros algoritmos similares es que si dos palabras tienden a aparecer en los mismos textos posiblemente sea porque se refieren al mismo tema. De esta manera, por coocurrencia de palabras, el algoritmo es capaz de inferir los temas o *topics*.

Estos métodos no supervisados son eficientes para tratar con grandes volúmenes de información textual y son capaces de extraer regularidades de ellos, sin embargo, presentan también algunos problemas. Entre otras cosas, por ejemplo, a veces cometen errores impredecibles porque comienza los procesos de inferencia con decisiones aleatorias que, si bien luego son corregidas mediante procesos iterativos, siempre dejan algún tipo de error. Así ocurre con *topic modeling* y otros algoritmos de *text mining*. Además, no

[7] Cfr. David M. Blei, «Probabilistic Topic Models», en *Communications of the ACM*, 55:4 (2012), pp. 77-84.

[8] Entre otros ejemplos de aplicación de *Topic Modeling* al análisis del texto literario, véase Matthew L. Jockers y David Mimno, «Significant Themes in 19th-Century Literature», en *Poetics*, 41 (2013); Christof Schöch, «Topic modeling genre: an exploration of french classical and enlightenment drama», en *Digital Humanities Quarterly*, 2017, DOI: 10.5281/zenodo.166356; Navarro Colorado, «On Poetic Topic Modeling...», *art. cit.*

todos los análisis lingüísticos y/o literarios de interés para los estudios literarios se pueden modelar con modelos de aprendizaje no supervisado. Estos métodos no supervisados son apropiados para extraer información, detectar regularidades en grandes colecciones de textos o calcular similitudes, pero no son del todo apropiados para aquellos análisis que utilizan taxonomías y clasificaciones previas[9].

Los métodos supervisados, dado que parten de un análisis humano, se pueden adaptar mejor al tipo de análisis profundo que precisan los estudios literarios. Actualmente, sin embargo, el mayor problema para desarrollar sistema de análisis del texto literario con métodos supervisados es precisamente la falta de corpus literarios anotados por expertos (los llamados *Gold Standards*). Para poder realizar sistemas computacionales que ayuden a analizar en profundidad grandes cantidades de texto literario es por tanto necesario disponer de estos corpus literarios anotados. Mediante la anotación, los aspectos específicos del análisis literario (bien de tipo temático, estilístico, lingüístico, etc.) quedan representados de manera explícita y formal, de tal manera que el sistema puede aprender cómo está analizado el texto y analizar así nuevos textos (literarios) con esta información.

Con el objetivo de realizar un análisis al tiempo distante (de gran cantidad de textos) y profundo (sobre rasgos lingüístico-literarios más o menos implícitos en los textos) de la lírica del Siglo de Oro, en este trabajo se propone la creación de un corpus general de referencia de este tipo de poesía. La propuesta no es tanto compilar un corpus único representativo sino un amplio corpus de referencia general que pueda dar soporte a cualquier tipo de estudio. Más allá de la selección de poemas, lo característico del corpus es que los poemas sean anotados con información relevante para los estudios literarios. Un corpus de estas características es oportuno tanto para el análisis del corpus en sí mismo, como para el desarrollo de sistemas automáticos de análisis textual específicos para el texto poético con métodos supervisados. Es, en definitiva, un recurso necesario para poder realizar análisis poéticos distantes y profundos.

Existen diversos proyectos para la creación de repertorios digitales literarios y poéticos en el ámbito hispánico. Entre ellos podemos destacar proyectos como ReMetCa[10] y PoeMetCa[11], ambos centrados en aspectos métricos de

[9] Sobre la aplicación de métodos supervisados y no supervisados a problemas de investigación literaria, así como el uso de clasificaciones previas, véanse las interesantes reflexiones y propuestas que T. Underwood presenta en *Distant Horizons. Digital evidence and literary change*, University of Chicago Press, 2019.

[10] Enlace: <http://www.remetca.uned.es/index.php?lang=es> [consulta: 06/12/2018].

[11] Enlace: <http://poemetca.linhd.uned.es/> [consulta: 06/12/2018].

la poesía de cancionero; la *Base de datos da Lírica Profana Galego-Portuguesa* (MedDB)[12]; la edición lírica y musical de las *Cantigas de Santa María de Alfonso X el Sabio*[13]; el *Corpus del Troubadours*[14] o el *Corpus de Sonetos del Siglo de Oro*[15]. La mayoría de estos proyectos están centrados en cuestiones de localización, catalogación, digitalización y edición crítica de los textos. Salvo la anotación métrica presente en algunos de ellos, ninguno de estos proyectos incorpora en el corpus la anotación de rasgos lingüísticos y/o literarios profundos como los que se proponen en este trabajo. Así, un corpus general de referencia como el aquí presentado viene a ser un paso más allá en la creación de recursos digitales para el análisis en profundidad del texto literario.

En la siguiente sección se expondrán y justificarán primero las especificaciones mínimas que, a mi juicio, un *corpus* de estas características debe tener. Se comentará sobre todo el tipo de información que debe ser anotada. A partir de estas especificaciones, para comprobar su validez y determinar los problemas que puedan surgir, se presentará en la siguiente sección la creación, anotación y validación de un *corpus* piloto. Este consta actualmente de 51.223 versos anotados con información estructural, métrica y categorial; de los cuales más de 5000 han sido revisados y corregidos por expertos. Finalmente se expondrán algunas estadísticas del corpus. El corpus piloto está publicado *on-line* y a disposición de la comunidad científica.

2. Especificaciones mínimas para un *corpus* general de referencia de la lírica castellana

En esta sección se expondrán los aspectos mínimos que deben estar presentes a la hora de crear un corpus general de referencia de la lírica castellana del Siglo de Oro. Los problemas tanto técnicos como histórico-literarios que un proyecto de este tipo puede presentar son bastantes y variados. No se pretende aquí dar cuenta de todos ellos, sino especificar únicamente los aspectos que considero mínimos o básicos teniendo en cuenta tanto, por un lado, el carácter general del corpus (sin un objetivo de análisis concreto) como, por otro, la situación actual en digitalización de textos y en procesamiento del lenguaje natural.

[12] Enlace: <https://www.cirp.gal/pls/bdo2/f?p=MEDDB3:2:3569717806443323032> [consulta: 06/12/2018].

[13] Enlace: <http://www.cantigasdesantamaria.com/> [consulta: 06/12/2018].

[14] Enlace: <https://trobadors.iec.cat/> [consulta: 06/12/2018].

[15] Enlace: <https://github.com/bncolorado/CorpusSonetosSigloDeOro> [consulta: 06/12/2018].

En concreto, los aspectos mínimos que un corpus así debe tener son tres:
- el texto, que debe ser estar fijado con criterios filológicos a partir de los principales testimonios conocidos;
- los metadatos, que deben situar cada poema en su contexto bibliográfico (autoría, ediciones previas, fechas de composición y publicación, etc.), y
- el enriquecimiento del texto mediante la anotación explícita de información lingüística y literaria[16].

Evidentemente, el primer aspecto básico que necesita un corpus digital de referencia es un texto de calidad. No solo es necesario disponer de los poemas en soporte digital, sino que el texto debe haber sido fijado con criterio y rigor filológico. Lo ideal es disponer de ediciones críticas digitales que den cuenta de los diferentes testimonios que nos han llegado de cada poema, las principales variantes de los textos y cuál debería ser la lectura correcta o apropiada a juicio del crítico. Si no fuera posible disponer de ediciones críticas completas con todo el estudio de testimonios y variantes (y a día de hoy no es posible), un corpus de referencia debe contar al menos con una edición del texto rigurosa que fije una lectura válida. Por desgracia, si bien muchas de las obras líricas del Siglo de Oro cuentan con buenas ediciones modernas en papel, no hay apenas, salvo algunas excepciones[17], ediciones críticas digitales. Las ediciones digitales de poemas del Siglo de Oro actuales son ediciones divulgativas con más errores a veces de lo deseado. La creación de buenas ediciones críticas digitales de las grandes obras de la literatura española es una de las tareas más urgentes hoy para los estudios literarios digitales.

Además de la fijación del texto, un corpus de poesía lírica general y de referencia debe situar también cada poema en su contexto bibliográfico. Quiero decir con esto que el corpus debe aportar información sobre las principales

[16] Sobre el enriquecimiento del texto literario digital con información lingüística y literaria, véase el concepto de «*smart big data*» propuesto por Christof Schöch, «Big? Smart? Clean? Messy? Data in the Humanities», en *Journal of Digital Humanities*, 22 (2013). Enlace: <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/> [consulta: 09/10/2019].

[17] A destacar, primero, la magnífica edición crítica digital de las *Soledades* de Luis de Góngora realizada por Antonio Rojas Castro cfr. «La edición crítica digital y la codificación TEI. Preliminares para una nueva edición de las Soledades de Luis de Góngora», en *Revista de Humanidades Digitales*, 1 (2017), pp. 4-19. Enlace: <http://www.soledadesediciondigital.com/> [consulta: 06/12/2018]., también son destacables las ediciones críticas digitales de las comedias de Lope de Vega realizadas por el grupo Prolope, enlace: <http://prolope.uab.cat> [consulta: 06/12/2018], las ediciones de Clásicos Hispánicos, enlace: <http://www.clasicoshispanicos.com/> [consulta: 06/12/2018] y las ediciones digitales realizadas por el grupo de investigación GRISO, enlace: <https://www.unav.edu/web/griso> [consulta: 06/12/2018]. Para una lista de fuentes textuales digitales en español, véase José Calvo Tello, *Atlas de Datos*, Würzburg, Universität Würzburg, 2016. Enlace: <https://github.com/morethanbooks/atlas-de-datos> [consulta: 06/12/2018].

ediciones de cada poema y a partir de cuál o cuáles se ha creado la edición digital, así como otros datos generales como autoría, fechas de publicación, codificación, modernización del texto, idiomas y demás datos bibliográficos[18].

En este punto se presenta un segundo problema, no menos complejo que la edición crítica, como es la autoría de muchos poemas del Siglo de Oro. Efectivamente, si bien la obra lírica de muchos autores llegó a ser impresa en libro durante el Siglo de Oro (empezando por la obra del propio Garcilaso), el principal medio de transmisión de aquella época fue el pliego suelto y el manuscrito[19]. Esto ha generado graves problemas de autoría para muchos poemas pues se dispone de testimonios y versiones diferentes asignados a diversos autores[20].

Ambos aspectos son cada uno en sí mismo un problema de investigación sobre los que existe abundante bibliografía y no vamos a entrar aquí. No es el objetivo de este artículo profundizar en los problemas crítico-textuales de la lírica del Siglo de Oro. Es en el tercer aspecto en el que nos vamos a centrar en este artículo: la necesidad de anotar el corpus de poesía con información lingüística y literaria, y determinar qué tipo de información debe ser anotada y cómo. Sirvan estas notas para dar cuenta del problema.

El objeto de este trabajo es por tanto dar un paso más y plantear la necesidad de disponer no solo de un texto literario fijado y documentado, sino también disponer de un texto literario enriquecido con información implícita (lingüística y literaria) para poder realizar así análisis al tiempo distantes (el análisis de gran cantidad de poemas) y profundos (el análisis de rasgos lingüístico-literarios implícitos).

El tipo de información que se debe anotar de manera explícita en un corpus digital depende sobre todo del análisis que se quiera realizar con ese corpus. Ahora bien, en el caso de un corpus general de referencia como el aquí planteado, la anotación no se crea con un objetivo analítico concreto. Por ello se debe anotar aquella información básica que sea útil para la mayor cantidad de estudios posible pero sin considerar ninguno análisis en concreto. De la misma manera, la anotación no debe responder a ninguna teoría concreta o

[18] Cfr. Lou Burnard, «Metadata for corpus work», en *Developing Linguistic Corpora: a Guide to Good Practice*, ed. de Martin Wynne, Oxford, Oxbow Books, 2005, pp. 30-46. Enlace: <http://ota.ox.ac.uk/documents/creating/dlc/> [consulta: 06/12/2018].

[19] Cfr. Alberto Blecua, *Manual de crítica textual*, Madrid, Castalia, 1983.

[20] Los mayores problemas de autoría se presentan con los romances, pues muchos de ellos son de inspiración popular. Sobre los problemas de autoría del romancero nuevo y su tratamiento formal basado en TEI, véase Raquel López Sánchez y Borja Navarro Colorado, «Propuesta teórico y metodológica para el desarrollo de un corpus digital representativo del romancero nuevo», en *Corpus y bases de datos para la investigación en literatura*, ed. de Rebeca Lázaro Niso, Logroño, Fundación San Millán de la Cogolla, 2017.

método específico. Además, dado el amplio tamaño del corpus aquí planteado, el tipo de información a anotar debe estar en consonancia con la situación actual de las técnicas de procesamiento del lenguaje natural[21]. Junto a un lenguaje formal de marcado de uso común (XML) y un estándar para la anotación de textos que permiten la representación de prácticamente cualquier fenómeno lingüístico o literario (la *Text Encoding Initiative* o TEI)[22], con los últimos avances en procesamiento del lenguaje natural se dispone ya de herramientas robustas para analizar de manera automática textos a diferentes niveles de descripción lingüística: léxico, morfológico, sintáctico, semántico e incluso textual y pragmático[23].

Sin embargo, la aplicación de las actuales herramientas de procesamiento del lenguaje natural al texto literario presenta serios problemas dado que no están preparados para ese tipo de texto[24]. Destacaré dos problemas principales. En primer lugar, el texto literario en general y el poético en particular exprime al máximo los recursos expresivos del lenguaje. Por ello los textos literarios suelen ser los más complejos de analizar tanto a nivel léxico como a nivel sintáctico y semántico. A la ambigüedad propia del idioma se le une la ambigüedad artística provocada por el propio autor y característica del género[25]. En segundo lugar, los sistemas de procesamiento del lenguaje natural suelen estar entrenados y preparados para analizar texto moderno y estándar (normalmente texto periodístico). Para tratar con textos más específicos, como textos médicos o textos breves de redes sociales, se deben adaptar las herramientas existentes. Esta adaptación es mucho más compleja para el texto literario, pues éste no es ni texto estándar ni (en la mayoría de las ocasiones) moderno. Todo lo contrario, la desautomatización de la lírica

[21] Cfr. James Pustejovsky y Amber Stubbs, *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*, Sebastopol, O'Reilly, 2012.

[22] Cfr. Enlace: <http://www.tei-c.org/> [consulta: 06/12/2018]. Si bien es cierto, como luego se mostrará, que el procesamiento del lenguaje natural tiene sus estándares propios para la representación de información lingüística.

[23] Para una introducción al procesamiento del lenguaje natural, véase Steven Bird, Ewan Klein y Edward Loper, *Natural language processing in Python*, Sebastopol, O'Reilly, 2009 Enlace: <http://www.nltk.org/book> [consulta: 06/12/2018]; Daniel Jurafski y James H. Martin, *Speech and Language Processing*, New Jersey, Prentice Hall, 2008; Christopher D. Manning y Hinri Schütze, *Foundations of Statistical Natural Language Processing*, *ob. cit.* Para el español, véase la revista *Procesamiento del lenguaje natural*, enlace: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/242> [consulta: 06/12/2018].

[24] Sobre la aplicación de técnicas de procesamiento del lenguaje natural al texto literario, véanse las actas del *Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Enlace: <https://sighum.wordpress.com/events/latech-clfl-2019/> [consulta: 06/12/2018].

[25] Cfr. Adam Hammond, Julian Brooke y Graeme Hirst, «A Tale of Two Cultures: Bringing Literary Analysis and Computational Linguistics Together», en *Workshop on Computational Linguistics for Literature*, Atlanta, 2013.

conlleva un estilo muy marcado y diferente del estándar, a lo que se suma la antigüedad de los textos.

Aunque hoy día no haya sistemas de procesamiento del lenguaje natural adaptados al texto literario[26], una forma viable de crear un corpus literario anotado es haciendo un primer análisis automático con las herramientas disponibles y luego corrigiendo la anotación a mano (creación de un *Gold Standard*). Una vez corregido el corpus, el sistema se puede entrenar de nuevo con la anotación correcta y así ir progresivamente mejorando el corpus y adaptando el sistema al texto literario. Este proceso semiautomático es el método estándar para anotar cualquier tipo de texto[27].

Así, por tanto, a partir de la situación actual de la técnica en procesamiento del lenguaje natural, consideramos que un corpus general de referencia de la lírica del Siglo de Oro debe disponer de información explícita a nivel sobre todo métrico-rítmico, morfo-léxico y textual. Junto a ello, en una versión más avanzada de la anotación, ésta podría completarse también con información sintáctica y semántica.

En el nivel métrico-rítmico existen ya corpus digitales de poesía española que tienen marcada información sobre la rima y la métrica como los proyecto ReMetCa[28] y PoeMetCa[29] o el *Corpus de Sonetos del Siglo de Oro*[30]. En este último caso se propone un modelo formal de representación métrica, entendida como secuencia de sílabas tónicas y átonas. Se dispone también de herramientas de procesamiento del lenguaje natural para la anotación automática de patrones métricos[31]. Si bien no es un tema cerrado, la métrica es el aspecto rítmico-poético que más atención ha suscitado en las Humanidades Digitales, quizá por su fácil formalización. Es un tipo de información básica que un corpus de poesía de referencia debe incluir.

Más allá de la métrica hay otros aspectos rítmicos de los que también se debería dar cuenta en la anotación del corpus, entre ellos el encabalgamiento[32], las pausas potenciales, líneas de entonación y otros aspectos prosódicos.

[26] Precisamente para adaptar estos sistemas al texto literario es necesario un corpus anotado de propósito general como el aquí planteado.

[27] Cfr. James Pustejovsky y Amber Stubbs, *Natural Language Annotation for Machine Learning*, *ob. cit*.

[28] Enlace: <http://www.remetca.uned.es/index.php?lang=es> [consulta: 06/12/2018].

[29] Enlace: <http://poemetca.linhd.uned.es/> [consulta: 06/12/2018].

[30] Cfr. Borja Navarro Colorado, María Ribes Lafoz y Noelia Sánchez, «Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation», en *LREC 2016, Tenth International Conference on Language Resources and Evaluation* Portoroz (Eslovenia), 2016 pp. 4360-4364. Enlace: <https://github.com/bncolorado/CorpusSonetosSigloDeOro> [consulta: 06/12/2018].

[31] Cfr. Borja Navarro Colorado, «A Metrical Scansion System for Fixed-Metre Spanish Poetry», en *Digital Scholarship in the Humanities*, 33:1 (2018), pp 112–127.

[32] Cfr. Pablo Ruiz Fabo, Clara I Martínez Cantón, Thierry Poibeau y Elena González-Blanco, «Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets», en *Joint SIGHUM Workshop on*

Estos aspectos, sin embargo, son más complejos y requieren procesos automáticos específicos. En todo caso, siendo la métrica y sobre todo la rítmica uno de los aspectos definidores de la lírica, un corpus general y referencia debe tener esta información marcada explícitamente.

El nivel léxico-morfológico es quizá el nivel de descripción lingüística donde más avances ha cosechado el procesamiento del lenguaje natural con los llamados «*PoS-taggers*» (*Part-of-Speech taggers*). Estas herramientas toman un texto sencillo y determinan, para cada palabra, su lema (la forma no marcada), su categoría gramatical e información morfológica relevante. Entre los *PoS-taggers* disponibles para español destacan *Freeling*[33], *CoreNLP* entrenado para español[34], *Maltparser* entrenado para español[35] o *CLiPS Patterns*[36].

El principal problema que deben resolver estos sistemas es la ambigüedad categorial: determinar la categoría gramatical de una palabra que, por su forma, podría pertenecer a dos o más categorías. Así, por ejemplo, «bajo» puede ser una preposición, un adjetivo o un nombre, o «cura» puede ser un nombre o un verbo. Esta ambigüedad se resuelve a partir de la información que aporta el contexto de cada palabra. En general estos sistemas funcionan con una precisión de análisis superior al 90% en texto moderno y estándar.

Su aplicación al texto literario presenta los problemas antes comentados: a la ambigüedad propia del texto literario se le une el uso de palabras en contextos que no son los habituales. Así, al estar entrenados para analizar textos modernos, suelen analizar incorrectamente palabras antiguas, formas arcaicas y en general palabras de uso poco común. Por ejemplo, al analizar poesía del Siglo de Oro con *Freeling* se ha detectado que suele analizar la palabra «do» como nombre (la nota do), cuando en este tipo de poesía lo más común es que «do» sea una contracción del adverbio «donde». Dado que esta contracción no es común en la lengua moderna estándar, el sistema tiende a analizarla de manera incorrecta[37].

A pesar de estos errores, la información categorial (lema, categoría gramatical e información morfológica) debe estar presente en un amplio corpus general de referencia de la lírica del Siglo de Oro. Por un lado, es información

Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 2017, pp. 27-32. Enlace: <http://aclweb.org/anthology/W17-2204> [consulta: 06/12/2018].

[33] Cfr. Lluis Padró y E. Stanilovsky, «FreeLing 3.0: Towards Wider Multilinguality», en *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, 2012. Enlace: <http://nlp.lsi.upc.edu/freeling/node/1> [consulta: 06/12/2018].

[34] Enlace: <https://nlp.stanford.edu/software/spanish-faq.shtml> [consulta: 02/12/2018].

[35] Enlace: <http://www.iula.upf.edu/recurs01_mpars_uk.htm> [consulta: 02/12/2018].

[36] Enlace: <https://www.clips.uantwerpen.be/pages/pattern-es> [consulta: 02/12/2018].

[37] Cfr. Borja Navarro Colorado, «A Metrical Scansion System for Fixed-Metre Spanish Poetry», *art. cit.*

lingüística básica y relevante para variedad de estudios y análisis estilísticos y lingüísticos; y por otro es viable anotar gran cantidad de textos literario con los sistemas actuales de procesamiento del lenguaje natural. Habrá, claro, un porcentaje de error que deberá ser revisado y corregido a mano hasta obtener un *Gold Standard*, una muestra del corpus con la anotación correcta.

El tercer nivel que consideramos básico para el desarrollo de un corpus general de referencia de la lírica del Siglo de Oro es el textual. En concreto, consideramos básico marcar de manera explícita el tipo de poema o, según el caso, el tipo de estrofa: soneto, lira, tercetos encadenados, romance, etc. Esta información es relevante porque muchos aspectos formales y semánticos dependen de ella. Así, de un romance, además de los versos octosílabos con rima asonante, esperamos una estructura más o menos narrativa.

No hay hoy día un sistema automático que clasifique poemas según su tipo o estrofa. Sin embargo, existen diversas técnicas de procesamiento del lenguaje natural para realizar clasificación textual. A partir de la anotación manual de una parte del corpus, se puede entrenar un sistema de clasificación textual que permita clasificar el resto del corpus, y así, con el proceso iterativo comentado ya, anotar el corpus entero. Para un clasificador de este tipo, una característica definitoria del tipo de poema es la información métrica marcada anteriormente.

Junto a estos tres niveles que consideramos básicos, hay dos aspectos más que deberían ser marcados en un corpus general de referencia para la lírica del Siglo de Oro pero que hoy son todavía problemas irresolubles para el procesamiento del lenguaje natural. Es precisamente por su especial complejidad por lo que considero que sería necesario disponer de un corpus de poesía con esta información marcada. Me refiero a información sintáctica por un lado y a información sobre usos metafóricos y simbólicos por otro.

El análisis sintáctico computacional es, junto al análisis categorial, una de las tareas más comunes en procesamiento del lenguaje natural. Es un análisis mucho más complejo que el análisis categorial, por lo que los resultados obtenidos actualmente no son tan precisos. Si bien los analizadores basados en dependencias (aquellos que determinan las relaciones sintácticas entre palabras) no son totalmente dependientes del orden de las palabras en la oración[38], en general a los analizadores sintácticos les resulta extremadamente complejo analizar oraciones desordenadas, es decir, el hipérbaton tan común en la lírica. Este fenómeno es muy específico de la poesía, por ello sería muy oportuno disponer de muestras de corpus poético anotadas con

[38] Cfr. Daniel Jurafski y James H. Martin, *Speech and Language Processing ob. cit.*, capítulo 3 de la 3ª edición *on-line*. Enlace: <https://web.stanford.edu/~jurafsky/slp3/13.pdf> [consulta: 07/12/2018].

árboles sintácticos para poder tratar el fenómeno del hipérbaton con modelos computacionales.

El análisis semántico es un campo del procesamiento del lenguaje natural bastante amplio en el que no voy a entrar aquí. Hay sin embargo un aspecto que sí considero debe ser tratado en un corpus general de poesía lírica. Me refiero a los usos metafóricos y simbólicos de las palabras. Los sistemas de procesamiento del lenguaje natural suelen centrarse en metáforas convencionales[39] que responden a procesos metafóricos cognitivos dentro del modelo de metáfora cognitiva desarrollada a partir de los trabajos de Lakoff y Johnson[40]. Son, por tanto, metáforas con un proceso de lexicalización más o menos desarrollado y sobre las que se han establecido correspondencias estables entre el término metafórico y el literal. La metáfora novedosa, más común en el texto poético, es un problema que apenas ha tenido interés en procesamiento del lenguaje natural[41]. Al igual que con el hipérbaton, es necesario disponer de *corpus* de poesía con los usos metafóricos marcados explícitamente para poder hacer un tratamiento computacional del fenómeno metafórico como elemento artístico de la poesía.

Por supuesto, hay muchísima más información literaria y lingüística que podría ser anotada en un corpus de referencia de la lírica del Siglo de Oro. Por ejemplo, una tarea muy común hoy día en procesamiento del lenguaje natural es la extracción de entidades nombradas en los textos como personas, lugares, organizaciones, productos, etc.[42]. Esta tarea sería interesante para los estudios distantes del Siglo de Oro si se pudieran anotar, por ejemplo, las entidades que hacen referencia a personajes mitológicos y/o históricos. Esta anotación permitiría hacer un análisis panorámico de los mitos comunes en la lírica del Siglo de Oro o de la relación de la poesía con hechos históricos. Esta información es, sin embargo, específica del periodo que, si bien sería muy provechosa su anotación, no se puede considerar anotación mínima y básica: por un lado, no es de interés para otros periodos de la lírica y, por otro, para su anotación se necesita información sobre categorías gramaticales que, esta sí, está propuesta como información básica. Por eso no la considero aquí

[39] Véase, entre otros, Ekaterina Shutova, «Models of Metaphor in NLP», en *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala (Suecia), 2010, pp. 688-697; Alice Deignan, *Metaphor and Corpus Linguistics*, Amsterdam, John Benjamins Publishing, 2005.

[40] Cfr. George Lakoff y Mark Johnson, *Metáforas de la vida cotidiana*, Madrid, Cátedra, 2017.

[41] Cfr. Vaibhav Kesarwani, Diana Inkpen, Stan Szpakowicz, Chris Tanasescu «Metaphor Detection in a Poetry Corpus», en *Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Vancouver, 2017, pp. 1-9.

[42] Cfr., entre otros, Bird *et al.*, *Natural Language Processing...*, *ob. cit.*, cap. 7; y Jurafski y Martin, *Speech and Language Processing*, *ob. cit.*, capítulo 17 de la 3ª edición *on-line*. Enlace: <https://web.stanford.edu/~jurafsky/slp3/17.pdf> [consulta: 07/12/2018].

para esta propuesta de mínimos. Quede en todo caso apuntada como trabajo futuro.

En esta sección, por tanto, se han especificado cinco elementos que a mi juicio deberían estar anotados en un corpus general de referencia de la poesía del Siglo de Oro. Los tres primeros los considero elementos básicos: el tipo de poema, la métrica y la categoría gramatical de cada palabra (junto al lema y demás información morfológica). Los otros dos componentes no son básicos pero sí necesarios para desarrollar sistemas de procesamiento del lenguaje natural específicos para poesía: las estructuras sintácticas de los poemas y los usos metafóricos y simbólicos de las palabras. En la siguiente sección se expondrá el desarrollo de un corpus piloto anotado con los tres aspectos básicos antes comentados.

3. Diseño y desarrollo de un corpus piloto de la lírica del Siglo de Oro de propósito general

Con el objetivo de crear un corpus general de referencia de la lírica del Siglo de Oro como el presentado en la sección anterior, se ha creado un corpus piloto de tamaño menor pero diverso en cuanto a la selección de textos, anotado a diferentes niveles y en parte revisado a mano. Con ello se ha estudiado la viabilidad de un corpus de estas características y de la información a anotar. En esta sección se presentará la selección de poemas del corpus piloto, la información anotada, su formalización y el proceso de anotación seguido.

3.1. Selección de textos

En la sección anterior se argumentó sobre la necesidad de marcar el tipo de poema o el modelo métrico-estrófico característico del poema. Para que el corpus piloto fuera diverso en cuanto a métrica y estructura, la selección de poemas se ha hecho en función de los principales tipos de poemas y modelos métrico-estróficos del Siglo de Oro. Se ha procurado que hubiera al menos un poema de cada tipo. De esta manera el corpus piloto está formado por diferentes muestras de tipos métricos y estróficos. En total en el corpus piloto hay 40 tipos diferentes de poemas o estrofas (según el caso)[43]. La Tabla 1 muestra aquellos con más cantidad de poemas en el corpus:

[43] Son los siguientes: canciones, coplas, coplas de pie quebrado, cuartetas, cuartetas asonantadas, cuartetos, cuarteto-lira, décimas, décimas espinela, églogas, elegías, endechas, epigramas, estancias, glosas, letrillas, liras, loas, madrigales, octavas, octavas reales, octavillas, octosílabos, odas, quintetos, quin-

TIPO	Poemas	TIPO	Poemas
Canción	40	Seguidilla	17
Liras	33	Tercetos	10
Romance	27	Madrigal	9
Silva	37	Glosas	8
Epigrama	24	Endechas	8
Octavas	25	Décimas	9
Tercetos encadenados	17	Quintillas	5
Sextetos lira	16	Villancicos	4
Égloga	15	Sonetos	4
Coplas	14	Cuartetas asonantadas	4
Octava real	17	Cuarteto-lira	4
Redondillas	14	Letrillas	3

Tabla 1. Tipos de estrofas y cantidad de poemas.

De esta manera quedan representados en el corpus si no todos sí los principales tipos de metros y de estrofas de la época. Aunque las cantidades no estén compensadas (ni por tipo ni por cantidad de versos), sí podemos considerar la selección como una muestra fiable a partir de la cual se pueden crear sub-corpus compensados para realizar análisis comparativos.

En total el corpus consta de 475 poemas, de los cuales 387 están clasificados según su modelo estrófico. Los 88 restantes aún no tienen la etiqueta de tipo de poema o bien generan dudas.

En la selección de autores se ha buscado también variedad, de tal manera que hubiera en el corpus poetas tanto del siglo xvi como del siglo xvii. Sin embargo, en la selección final ha primado el simple hecho de disponer de una versión digital de la poesía de estos autores. La mayoría de los poemas del corpus han sido extraídos de la *Biblioteca Virtual Miguel de Cervantes*[44], casi todos ya en formato procesable (HTML, que luego fue transformado a XML como luego se expondrá). La Tabla 2 muestra todos los autores del corpus junto a la cantidad de poemas y la cantidad de versos de cada uno.

tillas, redondillas, romances, romances heroico, seguidillas, septetos-lira, séptimas, sextetos lira, sextillas, sextinas, silvas, sonetos, tercetos, tercetos encadenados y villancicos.

[44] Enlace: <http://www.cervantesvirtual.com> [consulta: 06/12/2018].

Poeta	Siglo	Poemas	Versos
Alonso de Ercilla	xvi	10	5832
Antonio Enríquez Gómez	xvii	1	67
Pedro Calderón de la Barca	xvii	5	240
Cristóbal de Castillejo	xvi	14	199
Cristóbal de Virués	xvi	18	9312
Fernando de Herrera	xvi	7	972
Francisco de la Torre	xvi	22	3502
Francisco de Quevedo	xvii	24	1137
Fray Luis de León	xvi	83	5772
Garcilaso de la Vega	xvi	5	377
Gaspar Gil Polo	xvi	30	1607
Hernando de Acuña	xvi	17	2946
Jacinto Polo de Medina	xvii	61	2538
Juan Boscán	xvi	21	1500
Juan de la Cueva	xvi	4	3398
Lope de Vega	xvii	62	3145
Luis Barahona de Soto	xvi	1	326
Luis de Camoes	xvi	5	3694
Luis de Góngora	xvii	33	1731
Lupercio Leonardo de Argensola	xvii	4	370
Miguel de Cervantes	XVI	13	1070
San Juan de la Cruz	XVI	17	659
Santa Teresa de Jesús	XVI	5	95
Sor Juana Inés de la Cruz	XVII	13	734
Total	475	51223	

Tabla 2. Poetas del corpus, siglo y cantidad de poemas versos incluidos

El corpus piloto es por tanto variado porque no está centrado en ninguna escuela, movimiento o momento histórico concreto dentro del Siglo de Oro. Siguiendo la clasificación por épocas de José Manuel Blecua[45], el corpus incluye 15 poetas renacentistas (272 poemas, 41261 versos) y 8 poetas barrocos (203 poemas, 9962 versos).

El hecho de acudir a poemas ya digitalizados ha dejado fuera muchos autores poco conocidos. Así, de los 200 poetas incluidos por J. M. Blecua en su antología[46] (91 renacentistas y 109 barrocos), en el corpus solo hay 24. Un

[45] Cfr. José Manuel Blecua, *Poesía de la Edad de Oro I. Renacimiento*, Madrid, Castalia, 1984 y José Manuel Blecua, *Poesía de la Edad de Oro II. Barroco*, Madrid, Castalia, 1984.

[46] *Ibidem.*

corpus de referencia general deberá incluir a todos estos autores y otros[47], si bien para la mayoría de ellos aún no se dispone de una edición digital fiable.

En conclusión, la cantidad de autores del corpus piloto es pequeña en comparación con los testimonios de poemas de que se dispone hoy día. Sin embargo, para los propósitos del corpus piloto es suficiente: son más de 51000 versos de diferentes autores (tanto renacentistas como barrocos) y de diferentes formas métrico-estróficas (desde canciones y églogas hasta romances y coplas).

3.2. Enriquecimiento: anotación del corpus con información literaria y lingüística

Nuestro interés en combinar la perspectiva de análisis distante con el análisis en profundidad de fenómenos estilísticos y lingüísticos nos lleva, como se ha comentado antes, a la anotación del corpus de lírica del Siglo de Oro con información lingüística y literaria. De los diferentes niveles de anotación antes comentado, el corpus piloto ha sido anotado a nivel estructural, métrico y categorial.

3.2.1. Anotación estructural

La anotación estructural se ha centrado únicamente en la representación de los títulos de los poemas (normalmente el primer verso), el tipo de poema o de estrofa (si el poema es una sucesión de estrofas) y los versos. Para ello se han seguido las recomendaciones de la *Text Encoding Initiative*[48] y se ha utilizado las siguientes etiquetas:

- <head> y <title> para el título;
- <lg> para la estrofa, incluyendo *@type* para el tipo de poema o estrofa, además de *@resp* para dar cuenta del anotador responsable y *@cert* para indicar la certeza del anotador con la corrección de la anotación realizada (alta para casos claros, media o baja para casos dudosos); y
- <l> para el verso.

Los detalles de esta etiqueta de verso se indicarán más tarde. La figura 1 muestra un ejemplo de la anotación estructural:

[47] La Dra. María Ángeles Herrero, a partir de sus trabajos sobre poesía femenina de la edad moderna, está ya trabajando en la inclusión de mujeres poetas que escribieron y publicaron versos en esta época. Actualmente disponemos de una nómina de más de 30 mujeres poetas que se incluirán en el corpus en un futuro cercano. Cfr. María de los Ángeles Herrero, *L'univers literari de les escriptores valencianes dels segles xvi-xvii*, València, Institució Alfons el Magnànim, 2018.

[48] Enlace: <http://www.tei-c.org/> [consulta: 06/12/2018].

```
<text>
  <body>
    <head>
      <title>Al valeroso espíritu, ni suerte,</title>
    </head>
    <lg type="Octava_real" resp="maherrero" cert="high">
```

Figura 1. Muestra de la anotación estructural

Toda la anotación estructural se ha realizado automáticamente a partir o bien del texto sencillo o bien del texto en HTML. Durante la recopilación de los textos se fueron guardando los metadatos de cada poema y el tipo de poema o estrofa. En concreto, de cada poema se guardaba información sobre autor, primer verso, tipo de poema, fuente de la edición digital y *url*, y edición impresa original a partir de la cual se había creado la edición digital. Con todos estos datos se pudo generar automáticamente el XML con los datos del tipo de poema[49].

3.2.2. Anotación métrica

Para la anotación métrica se ha tomado como referencia la anotación del *Corpus de Sonetos del Siglo de Oro*[50]. Este corpus se centró solo en una forma poética muy concreta: el soneto. El corpus aquí planteado, como se ha mostrado, es mucho más variado en cuanto a tipos de metros, de poemas y de estrofas. Asumimos aquí el mismo concepto de patrón métrico: un patrón métrico es la secuencia de sílabas tónicas y átonas que se puede derivar de un verso. La unidad básica de representación es, por tanto, el verso y no se tienen en cuenta otros aspectos métricos como acentos secundarios, pausas potenciales, etc.

El modelo de representación del patrón métrico es, sin embargo, diferente. Mientras que el *Corpus de Sonetos* representa las sílabas tónicas y átonas con los símbolos «+» y «-» respectivamente, en este corpus piloto se indica explícitamente la posición de la sílaba tónica y la separación silábica. Así, la métrica del verso de *La Araucana* de A. de Ercilla «Al valeroso espíritu, ni suerte» quedaría representada así (Ejemplo 1):

(1) `<l met=»-|-|-|4|-|6|-|-|-|10|-|» n=»1»>`

[49] Dado que las ediciones digitales utilizadas han sido las de la Biblioteca Virtual Miguel de Cervantes, los metadatos solo dan información bibliográfica básica. El encabezado de cada poema es el mínimo que exige TEI con información sobre el proyecto, sobre la edición de referencia así como datos sobre codificación y anotación.

[50] Navarro Colorado *et al.*, «Metrical annotation of a large corpus...», *art. cit.*

El patrón métrico está asignado a la etiqueta @met. Las sílabas átonas se marcan con el símbolo «-». Las tónicas con la posición exacta de la sílaba. En este ejemplo, las sílabas tónicas son la 4ª, la 6ª y la 10ª. De esta manera se muestra de manera más evidente el tipo de endecasílabo. La separación silábica queda además representada explícitamente con la barra vertical «|». La razón de esta barra vertical tiene que ver con la sinalefa, como se expondrá en la sección siguiente.

La anotación métrica del corpus piloto, por tanto, si bien en su concepción es similar a la empleada en el *Corpus de Sonetos del Siglo de Oro*, es novedosa en cuanto a la representación al mostrar explícitamente la posición de las sílabas tónicas y la separación silábica.

3.2.3. Anotación categorial

Finalmente, el *corpus* piloto ha sido anotado con información morfo-léxica. En este nivel, de cada palabra se ha anotado su lema o forma no marcada y su categoría gramatical[51] junto a la información morfológica. La representación de la información categorial y morfológica se realiza mediante las etiquetas *Parole*[52], que son las etiquetas utilizadas por *Freeling* y otros *PoS-taggers* y están consideras el estándar *de facto* para el análisis categorial del castellano. La Figura 2 muestra el análisis categorial del mismo verso de A. de Ercilla:

```
<l met="-|-|-|4|-|6|-|-|-|10|-|" n="1">
  <span type="raw">Al valeroso espíritu, ni suerte,</span>
  <w lemma="a" type="SP">a</w>
  <w lemma="el" type="DA0MS0">el|</w>
  <w lemma="valeroso" type="AQ0MS00">va|le|ro|so</w>
  <w lemma="espíritu" type="NCMS000">es|pí|ri|tu|</w>
  <w lemma="ni" type="CC">ni|</w>
  <w lemma="suerte" type="NCFS000">suer|te|</w>
</l>
```

Figura 2. Análisis categorial.

[51] Las categorías gramaticales anotadas son: adjetivos, adverbios, artículos, determinantes, nombres, verbos, pronombres, conjunciones, numerales, interjecciones y preposiciones. Como es habitual en el análisis categorial, se marcan también abreviaturas y signos de puntuación.

[52] Enlace: <http://www.lsi.upc.es/~nlp/tools/parole-sp.html> [consulta: 06/12/2018].

Cada palabra está representada en una línea y marcada con la etiqueta <w>, que incluye el lema de la palabra en la etiqueta *@lemma* y la información categorial y morfológica en la etiqueta *@type*. Cada letra de la etiqueta categorial indica un tipo de información: la primera es la categoría gramatical y el resto la información morfológica. Las etiquetas del ejemplo de la Figura 2 se interpretan así:

SP = Preposición: «a»
DA0MS0 = Determinante (D) artículo (A) masculino (M) singular (S): «el»
AQ0MS00 = Adjetivo (A) calificativo (Q) masculino (M) singular (S): «valeroso»
NCMS000 = Nombre (N) común (N) masculino (M) singular (S): «espíritu»
CC = Conjunción (C) coordinada (C):«ni»
NCFS00 = Nombre (N) común (C) femenino (F) singular (S): «suerte»

Junto a la información categorial se ha introducido también la separación silábica de cada palabra a efectos métricos. Para ello se utiliza la barra vertical «|» en las mismas posiciones que en el patrón métrico. De esta manera hay una correspondencia absoluta entre las barras de las palabras y las del patrón métrico gracias a la cual es posible extraer de manera inequívoca cuáles son en concreto las sílabas de cada posición métrica.

Esta representación de la separación silábica podría parecer redundante. Se ha realizado así para dar respuesta a un problema de incompatibilidad provocado por la representación en un mismo formalismo de dos niveles de descripción lingüística.

Efectivamente, este formalismo está representando al mismo tiempo información métrica, cuya unidad mínima es la sílaba; e información morfológica, cuya unidad básica (a efectos de representación formal) es la palabra. La representación de ambos niveles resulta incompatible: la separación de palabras del análisis categorial entra en conflicto con la separación silábica de la métrica en todos los casos de sinalefa. Cuando se produce este fenómeno, dos sílabas de palabras diferentes se tratan como una única sílaba. Esto rompe los límites de la unidad palabra utilizado por el análisis categorial. Con cada sinalefa, por tanto, se produce una incompatibilidad entre los niveles representados, pues lo que a nivel métrico es una única sílaba y debe aparecer unido en la representación formal, a nivel léxico-categorial son dos palabras y debe aparecer separado.

Entre las recomendaciones TEI no encontramos ninguna forma óptima de representar a un mismo tiempo dos sílabas unidas por la métrica y separadas en el léxico. Cualquier opción con etiquetas XML implica un cruce de ramas,

lo cual genera un XML mal formado y por tanto inválido. La solución óptima que encontramos es utilizar la barra vertical para la separación silábica que en TEI aparece recomendada para la representación de la separación silábica en diccionarios[53]. Así, el final de cada palabra está representado con dos símbolos: la barra vertical «|» para indicar fin de sílaba y la etiqueta de cierre de palabra </w> para indicar fin de palabra. En aquellos casos donde se produce sinalefa no aparece barra vertical antes del cierre de palabra. Esto indica que la última sílaba de la palabra está unida por sinalefa a la primera sílaba de la palabra siguiente, como se puede ver en este ejemplo entre las palabras «valeroso espíritu» (ejemplo 2):

(2)
```
<w lemma=»valeroso» type=»AQ0MS00»>va|le|ro|so</w>
<w lemma=»espíritu» type=»NCMS000»>es|pí|ri|tu|</w>
```

De esta manera esta incompatibilidad queda representada de manera explícita y se evita el cruce de ramas.

En todo caso, la incompatibilidad entre niveles de anotación es un problema común. Hacer una única representación formal de diferentes niveles de descripción lingüística puede generar un formalismo difícil de analizar y procesar. Por ello consideramos que, para un corpus de referencia general, es mejor trabajar con diferentes versiones del corpus, una para cada tipo de información a representar: una versión para el nivel métrico, otro para el léxico-morfológico, otra para el nivel sintáctico, etc.

Al introducir la anotación categorial, la representación formal resultante es más compleja, por lo que a los anotadores les resulta difícil leer el verso original. Para facilitar la tarea de revisión y corrección se ha representado el verso sin anotación con la etiqueta ; la cual además facilita la recuperación automática del texto original sin ningún tipo de etiqueta.

3.2.4. Proceso de anotación

Toda esta anotación estructural, métrica y categorial se ha realizado en dos fases: una primera fase de anotación automática y una segunda fase de revisión, validación y (en su caso) corrección manual por parte de expertos en lengua y literatura.

Para la anotación automática se ha utilizado el sistema de escansión ADSO[54], que ha sido convenientemente ampliado y mejorado. Básicamente

[53] Cfr. Enlace: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html> [consulta: 06/12/2018].

[54] Cfr. Borja Navarro Colorado, «A Metrical Scansion System for Fixed-Metre Spanish Poetry», *art. cit.*, enlace: <https://github.com/bncolorado/adsoScansionSystem/tree/master/analysis> [consulta: 06/12/2018].

el sistema se ha mejorado en dos aspectos: en la salida del sistema y, sobre todo, en la capacidad para analizar cualquier tipo de verso.

En primer lugar, se ha modificado la salida del sistema para que muestre la separación silábica y la categoría gramatical de cada palabra según el formalismo especificado anteriormente. Esta información era ya utilizada por el sistema original para realizar el proceso de escansión, pero al finalizar el proceso se perdía al no quedar representada en la salida del sistema. Esta nueva versión simplemente genera una salida más rica con toda esta información. Como en el sistema original, el análisis categorial se realiza con *Freeling*[55].

En segundo lugar, el sistema se ha modificado para poder analizar la métrica de cualquier tipo de verso, no solo la del endecasílabo como en la versión original. Para ello se ha introducido un módulo nuevo que determina cuántas sílabas métricas debe tener el verso a partir de la cantidad de sílabas ortográficas; las posibles sinalefas, hiatos, sinéresis o diéresis, y la métrica de los versos anteriores. La explicación detallada, justificación y evaluación de este módulo excede los límites de este artículo.

Durante la segunda fase se ha revisado a mano el XML generado por el sistema automático y toda la anotación realizada. De los 475 poemas que forman el corpus piloto, se han revisado y, en su caso, corregido 5069 versos, un 10 % de la cantidad total de versos que componen el *corpus* piloto.

Los errores detectados en el texto no han sido corregidos. Se ha respetado en todo momento la edición de la *Biblioteca Virtual Miguel de Cervantes*. Los principales errores en el nivel métrico se han producido en la detección del tamaño del patrón métrico. Versos, por ejemplo, octosílabos con una posible sinalefa o sinéresis que el sistema erróneamente ha interpretado como heptasílabo y ha unido dos sílabas ortográficas en una única sílaba métrica. En cuanto al análisis categorial, los principales errores en la anotación automática se deben, como se comentó anteriormente, a formas y usos antiguos propios de los siglos xvi y xvii. Si bien los textos utilizados están modernizados, aún conservan contracciones, palabras y formas antiguas como por ejemplo «della», «hela» o «dixe». El uso de diéresis en algunas palabras para marcar bien la métrica ha generado problemas al analizador categorial, que no ha reconocido la palabra correctamente (por ejemplo, con «rüina»). Otros errores han sido producidos por la falta de información contextual: la propia ambigüedad del verso, la presencia de fuertes hipérbatos o la brevedad del contexto ha provocado que el sistema (y a veces hasta los propios anotadores) no dispongan de información suficiente para desambiguar las palabras.

[55] Cfr. Padró y Stanilovsky, «FreeLing 3.0...», *art. cit.*, enlace: <http://nlp.lsi.upc.edu/freeling/node/1> [consulta: 06/12/2018].

3.3. Datos generales del corpus piloto

Con toda esta información anotada se pueden mostrar ya algunos datos del corpus. A modo de ilustración se da cuenta aquí de las frecuencias máximas a nivel métrico y categorial.

Como era de esperar, los patrones métricos más frecuentes son los endecasílabos, los heptasílabos y los octosílabos. Entre los endecasílabos, los patrones más frecuentes son los acentuados en 2ª, 6ª y 10ª (3277 apariciones), en 3ª, 6ª y 10ª (3161 apariciones) y en 2ª, 4ª, 6ª y 10ª (2535 apariciones). Se echa en falta el patrón con sílabas tónicas en 2ª, 4ª, 8ª y 10ª, tan frecuente en los sonetos del Siglo de Oro[56]. Los heptasílabos más frecuentes son los patrones con sílaba tónica en las posiciones 2ª y 6ª (1120 apariciones) y en 2ª, 4ª y 6ª (849 apariciones). En cuanto a los octosílabos, el patrón más recurrente es el que presenta sílabas tónicas en 3ª y 7ª (695 apariciones) seguido de los patrones con tónicas en 4ª y 7ª (499 apariciones) y 2ª, 4ª y 7ª (473 apariciones).

En cuanto a las categorías gramaticales, teniendo en cuenta solo las categorías gramaticales abiertas, la que más se repite es el nombre común (61900 apariciones) seguido de verbo (49.352 apariciones) y del adjetivo calificativo (26511). El 10% del corpus piloto revisado a mano son en total 29885 etiquetas. La Tabla 3 muestra las categorías gramaticales más frecuentes de este 10 %:

Categoría gramatical	Ocurrencias	%
Nombre común	5868	19,6
Verbos	5158	17,2
Adjetivos	2404	8
Adverbios	1798	6
Conjunciones	2726	9,1
Pronombres	2202	7,3

Tabla 3. Frecuencias de categorías gramaticales

El nombre común es efectivamente la categoría abierta más frecuente con 5868 ocurrencias (19,6%). La presencia de verbos es de un 17,2% (5158) y de adjetivos de un 8% (2404). La forma verbal más utilizada son la tercera y la primera persona del singular del presente de indicativo.

La Tabla 4 muestra las frecuencias de categorías gramaticales abiertas por periodos. Mientras que en el Renacimiento la frecuencia de uso de nombres

[56] Cfr. Borja Navarro Colorado, «Hacia un análisis distante del endecasílabo áureo: patrones métricos, frecuencias y evolución histórica», en *Rhythmica. Revista Española de Métrica Comparada*, 14 (2016), DOI: https://doi.org/10.5944/rhythmica.18459.

y verbos es más o menos similar (sobre el 18%), en el Barroco se tiende a utilizar más nombres que verbos. Esta diferencia puede ser debida más a la diferencia en tamaño del corpus que a algún motivo literario.

Renacimiento	Ocurrencias	%	Barroco	Ocurrencias	%
Nombres	4781	18,9	Nombres	1087	23,7
Verbos	4540	17,9	Verbos	618	13,47
Adjetivos	1956	7,7	Adjetivos	448	9,7

Tabla 4. Frecuencias de categorías gramaticales por periodo

Estos datos tienen solo un propósito ilustrativo. Todavía no se pueden extraer de ellos conclusiones relevantes para los estudios literarios. Conforme se vayan revisando y validando más poemas se podrá ya entrenar y evaluar sistemas de procesamiento del lenguaje natural específicos para poesía y, en fin, se dispondrá de datos suficientes para un análisis no solo distante (que cubra gran cantidad de textos) sino también profundo (de rasgos lingüístico-literarios implícitos en el texto) de la lírica del Siglo de Oro.

3.4. Publicación y explotación del corpus piloto

Este corpus piloto es, por tanto, el primer paso ya dado para un análisis distante y profundo. El corpus piloto completo está publicado *on-line* y disponible para su descarga en el repositorio GitHub[57].

En el encabezado de cada poema se indica explícitamente si la anotación es la automática (y por tanto con algún posible error) o si ha sido ya revisada y validada a mano.

Ofrecer el corpus en XML siguiendo las recomendaciones de la TEI asegura la explotación del corpus con recurso estándar. Sin embargo, si bien estas tecnologías son de uso común dentro de las Humanidades Digitales, no son conocidas ni accesibles para los estudios literarios en general. Por ello es necesario ofrecer servicios web que permitan la explotación del corpus general de referencia tanto por parte de expertos en Humanidades Digital como por parte de estudiosos de la literatura en general[58]. Estos servicios están aún por definir y quedan como trabajo futuro.

[57] Enlace: <https://github.com/bncolorado/CorpusGeneralPoesiaLiricaCastellanaDelSigloDeOro> [consulta: 06/12/2018].

[58] Véase, por ejemplo, el buscador de versos por métrica desarrollado por la Biblioteca Virtual Miguel de Cervantes para el Corpus de Sonetos del Siglo de Oro. Enlace: <http://goldenage.cervantesvirtual.com/> [consulta: 06/12/2018].

4. Conclusiones

En este trabajo se argumenta a favor de combinar el análisis distante con el análisis profundo. Por análisis distante entendemos el análisis de gran cantidad de obras literarias para obtener análisis panorámicos de amplios periodos, normalmente aplicando técnicas computacionales de análisis textual. Estas técnicas, si bien cada día más avanzadas, no pueden aún dar cuenta de fenómenos implícitos del lenguaje (rasgos profundos), sobre todo en poesía.

Para poder combinar ambos análisis en el estudio de la lírica del Siglo de oro se plantea la necesidad de crear un corpus general de referencia. Como aspectos básicos este corpus debe disponer de un texto fijado de manera rigurosa, una correcta descripción de testimonio y ediciones de cada poema (incluyendo autoría, fechas, etc.) y una anotación correcta de aspectos lingüísticos y literarios generales. La anotación del corpus permite representar de manera explícita y formal rasgos lingüísticos y literarios implícitos que son relevantes para los estudios literarios. Es esta anotación la que permite combinar el análisis distante con el análisis profundo.

La anotación de un corpus puede ser muy variada dependiendo de los objetivos del análisis. Un corpus de referencia debe ser anotado con información general que sea útil para diversos tipos de análisis. Teniendo en cuenta el estado actual del procesamiento del lenguaje natural, se considera que un corpus de estas características debe estar anotado con información estructural (tipo de poema o de estrofa), métrica y categorial (lemas, categorías gramaticales e información morfológica de cada palabra). Además, se apunta la necesidad de dar cuenta de dos problemas complejos hoy para los sistemas de procesamiento del lenguaje natural: el hipérbaton y los usos metafóricos de las palabras.

En este sentido, un corpus de estas características es útil no solo como recurso de análisis en los estudios literarios, sino también como *Gold Standard* para desarrollar sistemas de procesamiento del lenguaje natural específicos para analizar la lengua literaria en general y la lírica en particular.

Para comprobar la viabilidad de este corpus general y determinar sus principales problemas, se ha creado un corpus piloto de poesía lírica del Siglo de Oro. El corpus consta de 51223 versos, abarcando 45 poemas de 21 poetas diferentes. La selección se realizó buscando muestras de diferentes modelos métrico-estrófico (romances, églogas, liras, etc.) a partir de textos ya digitalizados en la *Biblioteca Virtual Miguel de Cervantes*. Todo el corpus ha sido anotado en XML siguiendo las recomendaciones TEI a nivel estructural, métrico y categorial. El proceso de anotación ha sido semiautomático: tras

una primera anotación automática, ésta ha sido revisada, validada y, en su caso, corregida por expertos. Actualmente un 10% del corpus piloto ha sido ya revisado y corregido (5069 versos). El corpus se encuentra ya disponible para su descarga en la web, si bien la revisión y corrección manual continuará durante los próximos meses.

En conclusión, un corpus general de referencia de la lírica castellana del Siglo de Oro no solo es viable en estos momentos, sino totalmente necesario para poder desarrollar un análisis general y profundo de la poesía de los siglos xvi y xvii. Además de los diferentes aspectos a desarrollar antes comentados, desde la creación de ediciones críticas digitales hasta la anotación de hipérbatos o metáforas, es necesario también desarrollar servicios web que permitan la explotación del corpus por parte de toda la comunidad de investigadores interesados en la lírica del Siglo de Oro con independencia de sus conocimientos técnicos.

Recibido: 10/12/2018
Aceptado: 5/10/2019

Por un análisis distante y profundo: un corpus piloto de la poesía lírica castellana del Siglo de Oro

Resumen: En este trabajo se plantea la necesidad de combinar el análisis llamado «distante» (análisis panorámico de gran cantidad de texto literario) con el análisis profundo (análisis en detalle de diferentes aspectos lingüísticos o literarios). Para ello se propone la creación de amplios corpus literarios de referencia en los que, aprovechando los actuales avances en procesamiento del lenguaje natural, la información implícita del texto (tanto de tipo lingüístico como literario) esté marcada de manera explícita y formal. La propuesta se concreta en el diseño y desarrollo de un corpus piloto de la poesía lírica del Siglo de Oro que incluye poemas con diferentes modelos métrico-estróficos (sonetos, romances, liras, églogas, etc.) así como diversidad de autores. Actualmente consta de más de 52.000 versos anotados con información lingüística (palabras, lemas, categorías gramaticales y morfología) y literaria (tipo de poema o estrofa y métrica). Si bien la anotación general del corpus ha sido realizada de manera automática, un 10% de esa anotación (5069 versos) ha sido revisada, validada o, en su caso, corregida por expertos. Este 10%, en tanto que *Gold Standard*, es ya un primer paso tanto para el análisis distante y profundo de la poesía castellana como para el desarrollo de sistemas de procesamiento del lenguaje natural específicos para el texto literario y poético.

Palabras clave: Análisis distante. Poesía lírica. Siglo de Oro. Métrica. Procesamiento del Lenguaje Natural. Anotación de corpus.

Towards a distant and deep reading: a pilot corpus of Golden-Age Spanish poetry

Abstract: This paper shows the necessity of combine the distant reading of literary texts (panoramic analysis of a great amount of texts) with «deep» reading (close analysis in detail of implicit linguistic or literary aspects of texts). With this objective, the development of large annotated corpora of literary texts is proposed. Taking advantage of recent developments of Natural Language Processing, the linguistic and literary implicit information could be annotated semi-automatically. In order to show the viability of this proposal, a pilot corpus of Golden-Age Spanish poetry is presented. The corpus is made-up of different types of poems (sonnets, romances, eclogues, etc.) and several poets. Nowadays it has more than 52,000 lines annotated at metrical and morphological level: metrical patterns of each line, and the lemma, part of speech and morphological information of each word. The annotation was developed automatically. 5,069 lines has been revised manually and emended (if necessary). This Gold Standard is the first step both for a distant and deep literary analysis of Golden-Age Spanish poetry and for the development of poetry- specific models of Natural Language Processing.

Keywords: Distant reading. Poetry. Golden-Age. Meter. Natural Language Processing. Corpus annotation.