

# I. INTRODUCCIÓN. TEXTOS Y CALCULADORAS: LOS CORPUS DE LAS LENGUAS CLÁSICAS EN LA ERA DIGITAL

CRISTINA TUR Y BERTA GONZÁLEZ SAAVEDRA  
Universidad de Salamanca / Universidad Complutense de Madrid  
cristina.tur@usal.es / bertagon@ucm.es

La popularización del uso de ordenadores a partir de las últimas décadas del siglo xx y la difusión del acceso a internet han supuesto grandes cambios en la vida cotidiana: han variado las formas de comunicarse, de hacer transacciones, de realizar gestiones administrativas, de impartir o recibir clases. La Lingüística, así como el resto de las ciencias, no ha sido indiferente a las posibilidades que ofrece el desarrollo informático y se vale de ellas con frecuencia para tareas tan habituales como, por ejemplo, realizar una consulta en un diccionario en línea como el *Diccionario de la lengua española* o en los que ofrece el portal *Logeion* para el latín y el griego clásico o hacer búsquedas en el *Corpus de Referencia del Español Actual* (CREA), el *Corpus Diacrónico del Español* (CORDE) de la Real Academia Española, el corpus del *Packard Humanities Institute* y el *Thesaurus Linguae Latinae* para el latín clásico o, para el griego antiguo, el *Thesaurus Linguae Graecae*, entre otros.

El análisis lingüístico, pues, se sirve de las «nuevas tecnologías», que ya no son tan nuevas, y viceversa: la computación aprovecha en numerosas ocasiones estudios de lingüística para poder desarrollar herramientas como *chatbots*, traductores automáticos o cualquier otra aplicación en la que esté involucrado de una u otra forma el lenguaje humano. Fruto de la interacción entre Lingüística e Ingeniería informática ha surgido la Lingüística computacional,<sup>1</sup> que centra su atención, a grandes rasgos, en la comprensión y generación de lenguaje por parte de un ordenador, esto es, en lo que se conoce como Procesamiento del Lenguaje Natural (PLN a partir de ahora). Sin embargo, hacer que una potente calculadora, gobernada por las rígidas leyes de las matemáticas, sea capaz de manejar unos

---

<sup>1</sup> No puede establecerse un momento concreto en que se originó la Lingüística computacional, aunque está íntimamente relacionado con la investigación en el desarrollo de sistemas de traducción automática en los años cincuenta del siglo xx. Para una breve panorámica general de la Lingüística computacional, véase Kay (2003).

códigos repletos de excepciones a la norma, ironía, usos figurados, ambigüedades sintácticas y semánticas, etc., no es precisamente una tarea sencilla.

Una de las estrategias que se emplea para esta tarea es la de tomar un corpus de textos como modelo de la lengua en cuestión.<sup>2</sup> En este contexto se entiende por corpus un conjunto considerable de evidencias lingüísticas, normalmente de lenguaje en uso, ya sea oral o escrito (Sinclair 1996, p. 4; McEnery 2003). El tipo de corpus y origen de sus datos dependen del propósito de investigación para el que se ha confeccionado. De su creación, así como del desarrollo de herramientas para su consulta y utilización, se encarga la denominada Lingüística de corpus.<sup>3</sup>

## 1. LINGÜÍSTICA Y CORPUS

Se considera que el pionero de la Lingüística de corpus fue, precisamente, un latinista, el padre Busa, que, como él mismo relata (Busa 1980), a finales de los años cuarenta del siglo xx se puso en contacto con la empresa tecnológica IBM para realizar una concordancia de la obra de santo Tomás de Aquino. El *Index Thomisticus*, que fue el nombre que recibió este proyecto, empezó siendo una enorme colección de tarjetas perforadas, después, de cintas magnéticas y finalmente un CD-ROM, cuya información, en la actualidad, está accesible en internet. Posteriormente, y de forma paralela, se desarrollaron otros corpus (McEnery 2003), como el *Brown Corpus* (inglés) en la década de los cincuenta o, en los años noventa, el *British National Corpus* (inglés), el *Frantext* (francés), el CREA y el CORDE (español). El acceso computacional a los textos permite, sobre todo, hacer consultas muy rápidas y efectivas: al buscar un término concreto se puede conocer al instante cuántas veces aparece y en qué contextos se puede encontrar. Esta manera de extraer la información resulta, sin duda, más eficaz que la forma «tradicional» consistente en leer cada página y señalar manualmente cada aparición del fenómeno en cuestión.

Además, muchos de los textos que componen algunos corpus, como el *Brown Corpus* o el propio *Index Thomisticus*, están anotados lingüísticamente, es decir, cada una de sus palabras lleva asociada unas etiquetas (*tags*) que contienen información sobre sus características lingüísticas, normalmente morfológicas, en ocasiones también sintácticas y, en menor medida, semánticas y pragmáticas. De

---

<sup>2</sup> Existen otras aproximaciones al PLN que no requieren el uso de corpus, como las gramáticas libres de contexto, que se basan en reglas de formación y lexicones, o los autómatas de estados finitos y las expresiones regulares, que simbolizan patrones determinados mediante fórmulas. Para más información sobre ellas, consúltense Mitkov (2003) y Clark *et al.* (2013).

<sup>3</sup> Nótese que el concepto de *lingüística de corpus* es diferente del de *lenguas de corpus*, entendidas como aquellas de las que ya no hay hablantes nativos, pero de las que se conservan testimonios escritos. Sobre el estudio de la lingüística en las lenguas de corpus, véase Fernández Delgado *et al.* (1996).

este modo, se pueden depurar las consultas más fácilmente y buscar, por ejemplo, términos solo en cierto caso, género y número o palabras que desempeñan determinadas funciones sintácticas o semánticas.

El etiquetado de las palabras de un texto puede realizarse de diversas maneras, pero, en cualquier caso, ha de ser homogéneo en todos los textos que componen el corpus. Esto implica la creación de un estándar de anotación mediante el que se normalicen los análisis y que, además, pueda ser aplicado a otras colecciones de textos para, por ejemplo, poder construir corpus comparables, si se trata de distintos textos acerca de la misma materia, o paralelos, si se trata de los mismos textos en lenguas diferentes.<sup>4</sup> En este sentido, hay que resaltar el proyecto «Universal Dependencies», que nace en la segunda década de este siglo y que tiene como finalidad crear un estándar de anotación único que permita anotar lenguas de familias lingüísticas distintas. Para ello, tiene un conjunto de etiquetas de carácter universal, aunque también permite el etiquetado de fenómenos menos frecuentes tipológicamente mediante la creación de etiquetas, por lo que se da cabida a la singularidad de cada lengua.

La anotación de textos se puede orientar a diversos aspectos del análisis lingüístico. Por ejemplo, el *Brown Corpus*, uno de los primeros corpus anotados, registra las características morfológicas de cada palabra, pero *The Proposition Bank* (PropBank) se centra en los aspectos semánticos de las oraciones. Otros corpus analizados, también denominados bancos de datos (*banks*), incluyen el análisis de distintos planos lingüísticos, en especial el morfológico y el sintáctico, como el *Penn Treebank*. Gracias al etiquetado de análisis sintáctico, esto es, las relaciones entre las palabras, los textos se pueden visualizar en forma de árboles de dependencia, de ahí que los corpus etiquetados de esta forma reciban el nombre de *treebanks* ('bancos de árboles').

En lo que se refiere a la anotación sintáctica, hay dos modelos de análisis que condicionan la elaboración de los corpus: la gramática de dependencias y la de constituyentes. La primera se basa en el estructuralismo propugnado por Tesnière (1994), que establece un núcleo en cada oración, el verbo, del que dependen el resto de los elementos, formando una estructura de representación jerarquizada. La segunda, de corte generativista, se basa en una estructura arbórea bipartita, en la que cada rama se subdivide en sujeto y predicado, los dos elementos básicos de la oración (Matthews 1981, pp. 71-96). Como se puede suponer, estos dos enfoques sintácticos son muy diferentes; sin embargo, los corpus diseñados con un modelo cuentan con *scripts* de conversión para lograr, aunque no siempre con resultados satisfactorios, su compatibilidad con los corpus que siguen el otro.

---

<sup>4</sup> «Comparable corpora are corpora where a series of monolingual corpora are collected for a range of languages, preferably using the same sampling frame and with similar balance and representativeness, to enable the study of those languages in contrast. Parallel corpora take a slightly different approach to the study of languages in contrast, gathering a corpus in one language, and then translations of that corpus data into one or more languages» (McEnery 2003, p. 496).

El etiquetado de textos, sea cual sea el tipo de información que codifique, se puede realizar mediante un proceso semiautomático que consta de varias fases. La primera consiste en la *tokenización* del texto, es decir, en la división del texto en unidades menores: en primer lugar, las oraciones, delimitadas por una puntuación fuerte, como los puntos o los signos de exclamación o de interrogación. A su vez, las oraciones pueden dividirse en *tokens*, esto es, cualquier cadena de caracteres comprendida entre espacios en blanco.

Una vez que el texto está dividido en palabras y frases, se puede iniciar el proceso de análisis, que incluye la lematización, que es la asignación de cada palabra a un lema, y la adscripción a una categoría morfológica (sustantivo, verbo, adjetivo, etc.). También, según las características morfológicas de las palabras, se pueden establecer las relaciones de dependencia sintáctica, aunque suele haber un proceso posterior de revisión manual por parte de lingüistas. De igual manera, se pueden asignar determinadas características semánticas a algunos de los elementos.<sup>5</sup>

La tecnología que se emplea para etiquetar también puede ser variada, aunque el lenguaje de marcado más popular es el XML (*eXtensible Markup Language*), que, a pesar de no estar únicamente vinculado a la lingüística de corpus, constituye la base de, entre otros, el *Text Encoding Initiative* (TEI)<sup>6</sup> y del *Prague Markup Language* (PML),<sup>7</sup> desarrollado en el Instituto de Lingüística Formal y Aplicada de la Charles University de Praga, que es un ejemplo de estándar de anotación morfológica, sintáctica y semántica. En un principio, este último estándar se creó para la anotación de los textos en checo del *Prague Dependency Treebank* (Hajič *et al.* 2018), aunque se ha aplicado también a otras lenguas como, por ejemplo, el latín.

## 2. TREEBANKS PARA EL LATÍN Y EL GRIEGO ANTIGUO

Uno de los grandes retos de un *treebank* o un corpus de lenguas clásicas es que los textos que lo componen son obras literarias de gran envergadura de géneros como el teatro, la filosofía, la historia, etc. y esto hace que no sean frecuentes los estudios (inter)lingüísticos para establecer generalizaciones, sino que los usuarios de estos recursos están más interesados en aspectos lingüísticos de los propios textos incluidos en el corpus (González Saavedra y Passarotti 2018, p. 6).

<sup>5</sup> Para esto, se suelen emplear las llamadas *pipelines* ('tuberías'), en las que los procesos informáticos se encadenan de modo que el resultado de uno de ellos es la entrada del siguiente. Un ejemplo de *pipeline* con el que se pueden anotar textos según los criterios de *Universal Dependencies* es *UDPipe*, desarrollado por Instituto de Lingüística Formal y Aplicada de la Charles University de Praga (<https://ufal.mff.cuni.cz/udpipe/1>).

<sup>6</sup> <https://tei-c.org/>

<sup>7</sup> <https://ufal.mff.cuni.cz/pml>

Tal y como se ha señalado ya, uno de los eventos que marca el inicio de la Lingüística computacional es la intención de Roberto Busa de crear unas correspondencias del texto de Tomás de Aquino, así que podemos remontar la existencia de los corpus anotados para el latín al comienzo de la propia disciplina, pues el *Index Thomisticus* contiene, desde sus inicios, la lematización y la morfología de todas las palabras que componen la obra del santo. Este corpus, al principio del siglo XXI, volvió a marcar otro hito en los *treebanks* para las lenguas clásicas, pues en 2006 Marco Passarotti y Roberto Busa adaptaron el etiquetado sintáctico desarrollado en Praga para el *Prague Dependency Treebank* y crearon el *Index Thomisticus Treebank*, que consta de 450 000 elementos o nodos y de más de 26 000 frases anotadas sintácticamente, de las que unas 2 000 están anotadas con información semántica y pragmática (unos 28 000 nodos). Este sistema de anotación se estructura en cuatro niveles (*tokens*, morfología, sintaxis y semántica/pragmática), lo que facilita la extracción de la información. En cuanto a la sintaxis, sigue el modelo de la gramática de dependencias recogida en los principios de la Descripción Generativo-Funcional (Sgall *et al.* 1986).

Paralelamente, en 2008 nace el proyecto «Pragmatic Resources in Old Indo-European Languages» (PROIEL), con el fin de crear un corpus paralelo anotando las versiones del texto bíblico en varias lenguas indoeuropeas antiguas (Haug y Jøhndal 2008). Se trata de un corpus paralelo de lenguas que pertenecen a la misma familia lingüística y contiene anotación morfológica, sintáctica, semántica e, incluso, pragmática en un único nivel. La estructura de la sintaxis también sigue el modelo de la gramática de dependencias, aunque los principios de su análisis son propios, desarrollados para el propio proyecto.

A partir de los principios teóricos y de estructura de datos del *Index Thomisticus Treebank*, se desarrolla desde el 2006 el *Ancient Greek and Latin Dependency Treebank*, con los textos anotados morfológicamente del recurso *Perseus Digital Library* (Bamman y Crane 2011).<sup>8</sup>

Los *treebanks* aquí mencionados están incluidos en el repositorio general de *Universal Dependencies*, lo que quiere decir que se han establecido las reglas de conversión necesarias para que todos puedan ser consultados utilizando los mismos estándares de anotación y se puedan utilizar a modo de corpus comparables.

---

<sup>8</sup> Los autores incluidos en este repositorio son los siguientes: para griego, Esopo, Esquilo (siete tragedias), Ateneo (dos libros de *Deipnosophistas*), Diodoro Sículo (un libro de su *Biblioteca histórica*), Heródoto (un libro de su *Historia*), Hesíodo, Homero, Lisias (cuatro discursos), Polibio (un libro de las *Historias*), pseudo Apolodoro (un libro de su *Biblioteca*), el *Himno homérico a Deméter*, Sófocles (cinco tragedias) y Tucídides (partes del libro 1 de las *Historias*). Para latín, Augusto (*Res gestae* en preparación), César (parte de la *Guerra de las Galias*), Cicerón (parte de las *Catilinarias*), Jerónimo (*Vulgata*), Virgilio (parte de la *Eneida*), Ovidio (*Metamorfosis*), Petronio (parte del *Satiricón*), Fedro (parte de las *Fábulas*), Propercio (libro 1 de *Elegías*), Salustio (*Conjuración de Catilina*), Suetonio (parte de la *Vida de Augusto*) y Tácito (parte del libro 1 de las *Historias*).

Sin embargo, a la luz de lo expuesto hasta este momento se puede ver que el marco teórico y los ámbitos lingüísticos analizados en cada uno de los *treebanks* no son homogéneos y, aunque estén incluidos en *Universal Dependencies*, la información que se incluye en su estándar es solo morfológica y sintáctica. A esto hay que añadir que los textos no están anotados de forma completa cuando se trata de autores clásicos, y que, en el caso del latín y el griego no son comparables los autores de época clásica (en un sentido laxo) con santo Tomás desde un punto de vista ni cuantitativo ni del estadio de la lengua que representan.

Además, con la información sintáctica y semántica que contiene el *Index Thomisticus Treebank*, se han creado dos léxicos de verbos latinos en los que se recoge el funcionamiento sintáctico y semántico: *IT-vallex* (McGillivray y Passarotti 2009) y *LatinVallex* (González Saavedra y Passarotti 2016). Estas dos herramientas tienen la finalidad de recoger, a modo de diccionario, los principales verbos latinos con las estructuras mínimas de predicación que cada uno de ellos genera, esto es, sus marcos predicativos.

Por otra parte, el proyecto «Rección y complementación en griego y latín» (REGLA)<sup>9</sup> nace con la idea de crear un léxico de valencias de los verbos más frecuentes en griego antiguo y latín clásico, esto es, un catálogo lo más exhaustivo posible de los esquemas de complementación obligatoria o marcos predicativos de estos verbos. Para ello se diseñaron las bases de datos *REGLA-Latín* y *REGLA-Griego*, con un afán de estudio interlingüístico de las estructuras verbales desde la perspectiva de la Gramática Funcional (§ 4). Los datos consignados en estas bases de datos son de bastante calidad, ya que los análisis han sido llevados a cabo siempre por miembros del equipo de investigación, especialistas en lingüística y en griego y latín, pero la forma de organizar y almacenar estos datos ha planteado dos dificultades importantes: un análisis limitado a los constituyentes centrales de la predicación y la incapacidad para poner en relación oraciones principales y subordinadas.

### 3. COMREGLA

El proyecto COMREGLA «Compatibilidad de la base de datos REGLA con otros recursos digitales», desarrollado por investigadores de distintas universidades, surgió en 2018 con la intención de dar a estas cuestiones problemáticas una solución satisfactoria, así como de compatibilizar los datos disponibles en la base

---

<sup>9</sup> Este proyecto nació en el año 1992 por iniciativa de un grupo de investigadores de cuatro universidades españolas: Universidad Autónoma de Madrid, Universidad Complutense de Madrid, Universidad de Alcalá de Henares y Universidad de Santiago de Compostela, al que se fueron incorporando otras como la Universidad de Salamanca y la Universidad de Oviedo.

de datos REGLA con otras herramientas y recursos dedicados a las lenguas que nos ocupan.<sup>10</sup>

Para ello se consideró que era necesario hacer una migración de las dos bases de datos relacionales a una base de datos XML, llamada COMREGLA, que supuso un gran cambio estructural, puesto que la forma en que se almacena y estructura la información en las bases de datos relacionales es muy distinta a la de los sistemas de etiquetado XML.

Para su diseño, se tomó como modelo un *standard* XML ya existente para el análisis sintáctico y semántico necesario en la creación de *treebanks*, el *Prague Markup Language* (PML), un sistema de marcado desarrollado para el *Prague Dependency Treebank* y que ya ha sido aplicado al latín en el *Index Thomisticus Treebank*, entre otros recursos, tal y como ha se ha descrito.

El PML es, a grandes rasgos, un marcaje *stand-off*<sup>11</sup> que se articula en cuatro capas o niveles de análisis: *tokens*, morfología o nivel morfológico, análisis sintáctico o nivel analítico y análisis semántico-pragmático o nivel tectogramatical. No obstante, este sistema no resulta del todo compatible con el tipo de información que se almacena en las bases de datos relacionales, especialmente en el nivel sintáctico y semántico, en los que se siguen preceptos teóricos diferentes.<sup>12</sup> En este sentido, el PML resultaba insuficiente para reflejar determinada información sintáctica y semántica que se tiene en cuenta en REGLA, como, por ejemplo, las características semánticas de las predicaciones en su conjunto, sobre todo, cuando están subordinadas a una predicación principal.

Por tanto, se decidió que los elementos de la base de datos COMREGLA estuvieran anotados mediante un sistema propio de etiquetas XML, basado en buena medida en el PML, pero también en otros sistemas como PROIEL y ajustado lo más posible a los campos de las bases de datos relacionales de REGLA para poder mantener la precisión y rigurosidad del análisis sintáctico y semántico realizado hasta el momento. Mediante este sistema, no solo se superarían las dificultades antes mencionadas, sino que también se podría ofrecer la información a la comunidad científica en un formato compatible con los que se emplean en otros proyectos similares.

En las bases de datos de REGLA hay almacenados cuatro tipos de información lingüística: morfológica, sintáctica, semántica y léxica. En la nueva base de datos, esta información se ha redistribuido, como se observa en la siguiente tabla, en

<sup>10</sup> No hay que perder de vista que en 2018 nace en Milán el proyecto *Linking Latin* (LiLa), con el fin de crear una única plataforma que permita acceder de forma unificada a los distintos recursos dedicados a la lengua latina (Passarotti *et al.* 2019).

<sup>11</sup> El marcaje *stand-off* implica almacenar los análisis en varias capas de información. La primera capa es el texto plano, sin etiquetas, y sobre ella se van superponiendo las demás (información morfológica, después sintáctica y, por último, semántica). Para más información sobre este tipo de estructura y su aplicación en lingüística, véase, por ejemplo, Dipper (2005).

<sup>12</sup> Por ejemplo, el PML distingue entre argumentos y adjuntos obligatorios, mientras que en COMREGLA los adjuntos son por definición constituyentes opcionales del predicado.

dos niveles *stand-off*: *words*, en el que se recoge la forma, el lema y la información morfológica de cada palabra del texto, y *clauses*, donde se explicitan los rasgos léxicos de las unidades lingüísticas, las relaciones sintácticas y semánticas que se establecen entre ellas y las jerarquías de estructuras sintácticas de las que forman parte.

	WORDS	CLAUSES
Morfología	Forma y lema	–
	Características morfológicas	
Sintaxis	–	Palabras ( <i>words</i> ) < Predicaciones ( <i>clauses</i> ) < Oraciones ( <i>sentences</i> ) Relaciones sintácticas (dependencias, funciones sintácticas, etc.): <ul style="list-style-type: none"><li>entre las palabras de una predicación</li><li>entre las predicaciones que conforman una oración</li></ul>
Semántica	–	Características semánticas: <ul style="list-style-type: none"><li>de las relaciones (funciones semánticas, tipos de subordinación, etc.)</li><li>de las predicaciones (polaridad, diátesis, fuerza ilocutiva, control, aspecto léxico, etc.)</li></ul>
Léxico	–	Rasgos léxicos

TABLA 1. *Distribución de la información lingüística en los nuevos niveles (tomada de Garzón et al., 2023b, p. 80)*

Por otra parte, hay que tener en cuenta que los aspectos sintácticos que se recogen en la capa *clauses* parten de la división del texto en unidades, *tokens*, que son no solo palabras sino también otros tipos de unidades como la puntuación, números, etc., y que constituyen la forma más básica (*words*). Todas las unidades básicas que estén comprendidas entre puntuación fuerte conforman oraciones (*sentences*). La novedad de COMREGLA es que entre *words* y *sentences* sitúa una unidad intermedia, pero central en el análisis sintáctico y semántico de las bases de datos relacionales REGLA: las predicaciones (*clauses*). En este nivel, se recoge también la información semántica de las unidades, que definen el tipo de relación entre el verbo y sus elementos o los tipos de subordinación, así como otras características propias de la predicación como la polaridad, la diátesis, la fuerza ilocutiva o el aspecto léxico. En este nivel, además, se anota la información sobre el léxico de los elementos que funcionan como participantes en la oración.



#### 4. MARCO TEÓRICO: LA GRAMÁTICA FUNCIONAL

La base de datos REGLA, y también COMREGLA, se basa, desde sus inicios, en los fundamentos teóricos de la Gramática Funcional de S. Dik (1997). Este marco teórico se ha aplicado al estudio tanto del latín como del griego, como se puede observar en los trabajos desarrollados por H. Pinkster para el latín (2015, 2021) y los realizados por los miembros de REGLA tanto para el latín como para el griego (Baños *et al.* 2003; Torrego *et al.* 2007; Baños 2021; Jiménez López 2020b).

La Gramática Funcional adopta la predicación como unidad fundamental de análisis sintáctico-semántico. Según esta perspectiva teórica, toda predicación está formada por términos que designan entidades, esto es, objetos o conceptos del mundo real como un libro, la profesora o esa idea, o propiedades (atributos) de dichas entidades, como interesante o rápidamente. Pero estas entidades y atributos no aparecen de manera independiente, sino relacionadas a través de los predicados. Por ejemplo, en la predicación *Yo le he dado un libro a la profesora* hay tres entidades (*yo*, *libro* y *la profesora*) relacionadas mediante el predicado *he dado*; en cambio, en *esa idea es interesante*, el predicado relaciona una sola entidad (*esa idea*) con una propiedad (*interesante*). Las relaciones que se dan entre los elementos que integran las predicaciones son sintácticas y semánticas, es decir, los elementos que se relacionan con el predicado tienen determinados papeles sintácticos (sujeto, objeto, etc.) y semánticos (Agente, Paciente, Tiempo, etc.) en la predicación.

Los predicados son el núcleo semántico y sintáctico de esta estructura: semántico porque establece el tipo de relación que se da entre los diferentes términos; sintáctico porque determina cuántos elementos y de qué tipo son necesarios para que esté semánticamente completo. Por ejemplo, el predicado *dar* vincula necesariamente tres entidades: quien da, lo que se da y quien lo recibe. Estos elementos obligatorios se denominan argumentos y su número está definido por la valencia del predicado del que dependen. No obstante, el número no es el único aspecto con el que un predicado delimita sus argumentos, también determina sus rasgos léxicos. Por ejemplo, el primer argumento del predicado *leer* debe ser necesariamente una persona, pues leer es una actividad propia de los humanos (*mi padre lee el periódico*). Por tanto, este primer argumento ha de tener obligatoriamente el rasgo léxico [+ humano] o, de lo contrario, la predicación, aunque gramatical, no resultaría semánticamente aceptable (*una \*piedra lee el periódico*).<sup>13</sup> Este conjunto de predicado y argumentos se denomina predicación nuclear (Dik 1997, p. 291) y se puede formalizar mediante los llamados *Marcos Predictivos* (Dik 1997, pp. 78 y ss.; Villa 2003), que son los esquemas abstractos en

<sup>13</sup> Salvo, naturalmente, en casos en que se esté empleando el término *piedra* con un significado metafórico.

los que se da cuenta de las estructuras de complementación propias de cada predicado.

Junto al núcleo de la predicación pueden aparecer elementos no obligatorios que especifican características adicionales del evento. Estos constituyentes pueden aportar distinto tipo de información (Torrego y Villa 2021, pp. 37-39), o bien sobre el evento y sus argumentos (adjuntos), como τῇ δὲ τετάρτῃ ἡμέρᾳ ‘al cuarto día’ en (1) o *in oculis* ‘en la mirada / corazón’ en (2), o bien sobre el propio mensaje o los participantes en el acto de habla (disjuntos), caso de *ut dixi* ‘como dije’ también en (2). Una predicación nuclear y sus adjuntos forman una predicación extendida; si se le añaden los disjuntos, forman lo que aquí denominaremos, de manera general, simplemente predicación.

- (1) τῇ δὲ τετάρτῃ ἡμέρᾳ ἦκον οἱ τῶν πολεμίων ἱππεῖς  
(«Y al cuarto día apareció la caballería enemiga», Xen., *Ages.* 1.29)

Predicación nuclear: ἦκον οἱ τῶν πολεμίων ἱππεῖς

Predicación extendida: predicación nuclear + τῇ δὲ τετάρτῃ ἡμέρᾳ

- (2) *Te, ut dixi, fero in oculis*  
(«a ti, como te dije, te llevo en el corazón», Cic., *fam.* XVI 27.2)

Predicación nuclear: (*ego*) *te fero*

Predicación extendida: predicación nuclear + *in oculis*

Predicación: predicación extendida + *ut dixi*

Esta perspectiva funcionalista se ha enriquecido a lo largo de los años con aportaciones de otros marcos teóricos afines como la Gramática Cognitiva (Langacker 2008) o la Gramática de las Construcciones (Goldberg 1995), así como con otras teorías funcionalistas posteriores a las de Dik, como la Gramática del Papel y la Referencia (van Valin y LaPolla 1997) y la Gramática Funcional del Discurso (Hengeveld y Mackenzie 2008). Todas estas perspectivas comparten una visión de la lengua en la que priman la función comunicativa del lenguaje y el uso en contexto por encima de cuestiones puramente formales.

## 5. PROBLEMAS DE ANOTACIÓN: ¿QUÉ VAMOS A TRATAR EN ESTE LIBRO?

Como se acaba de comentar, la Gramática Funcional, en la que se enmarca teóricamente el presente libro, se articula en torno a las predicaciones, que pueden aparecer solas en una oración simple o relacionadas con otras formando una oración compleja, ya sea al mismo nivel (coordinación) o en jerarquía (subordinación). Sin embargo, en la mayoría de los *treebanks* mencionados con anterioridad, las unidades de análisis que se toman en consideración van directamente de la palabra (unidad mínima) a la oración (conjunto de palabras entre puntuación fuerte), sea esta simple o compleja. Esta división motiva que, en el capítulo II,

«Entre la oración y la palabra: la predicación como unidad de análisis», Cristina Tur argumente que esta perspectiva hace que el análisis semántico de las predicaciones sea menos preciso y completo, por lo que, al menos para el análisis lingüístico del latín y del griego antiguo, resulta conveniente la inclusión de una unidad intermedia que estructure el texto en predicaciones vinculadas entre sí tanto en relaciones de coordinación como de subordinación.

Precisamente, en el capítulo III, «Estructuras comparativas en griego antiguo y latín: tratamiento hacia un análisis unificado», Eveling Garzón aborda un tipo de subordinación que supone un reto para la anotación lingüística de cualquier corpus: la comparación. Prueba de ello es el análisis diverso que se hace de estructuras comparativas en apariencia diferentes, o incluso similares, en *treebanks* con anotación sintáctica y semántica dedicados específicamente a las lenguas clásicas. Con carácter general, las estructuras comparativas establecen la posición de una entidad (primer término o comparado) con respecto a otra (segundo término o estándar) en una escala graduable (cuantificador o marcador de grado) establecida sobre un determinado criterio (base o propiedad de comparación). Sin embargo, no siempre estos elementos aparecen explícitamente ni tampoco se desarrollan a través de términos independientes, sino que algunos de ellos pueden estar elididos por diversos motivos o estar integrados bajo un mismo término. En este capítulo se intenta dar una solución unificada para el análisis de estas estructuras.

Otro aspecto que se aborda en este libro se relaciona con el hecho de que las predicaciones nucleares, como se ha visto en § 4, pueden formalizarse a través de los Marcos Predicativos, lo que suele implicar una numeración de los argumentos del predicado. Así, algunas bases de datos, como ADESSE (*Alternancias de Diátesis y Esquemas Sintácticos-Semánticos del Español*) o REGLA tienen, por convención, un número concreto para los argumentos más frecuentes (Argumento 1, para el Agente; Argumento 2, para el Tema-Afectado-Efectuado; Argumento 3, para el Receptor / Beneficiario). Iván López, en el capítulo IV, «La numeración argumental y la diátesis: una propuesta de análisis», sostiene que este sistema, que resulta eficaz aunque sea convencional, puede, en ocasiones, plantear problemas cuando hay que analizar una predicación que presenta extensiones diatéticas diversas (pasivas y construcciones causativas fundamentalmente) y expone una solución que permite, por un lado, no constreñir el análisis argumental con esta numeración y, por otro, recuperar toda la información de los argumentos verbales.

Además de la diátesis, el aspecto léxico es otra cuestión que ha entrañado dificultades en el análisis, dado que se trata de una categoría en cuya definición intervienen otros constituyentes además de la forma verbal. Guillermo Salas, en el capítulo V, «La anotación del aspecto léxico en latín: más allá del predicado», argumenta que el modelo de anotación COMREGLA, al permitir anotar la información relativa al aspecto léxico en el nivel de la predicación, posibilita abordar satisfactoriamente la versatilidad aspectual de los predicados y reflejar de mane-

ra transparente el hecho de que se trata de una categoría que atañe a un ámbito que los trasciende.

Por otra parte, una de las ventajas de la anotación morfológica y sintáctica disponible para la mayoría de *treebanks* de lenguas clásicas es que permite no solo conocer cuál es la relación entre ambos niveles de análisis, sino también hacer estudios comparativos entre las lenguas, con el fin de determinar si las funciones sintácticas tienen los mismos rasgos morfológicos: objetos en acusativo, concordancia, etc. No obstante, incluir la notación semántica al análisis de los componentes de las predicaciones permite hacer comparaciones interlingüísticas (en el caso que nos compete, griego y latín), en las que se ve el vínculo entre morfología y semántica, lo que a su vez posibilita establecer la relación entre distintas funciones semánticas a partir de su coincidencia formal. Todo esto promueve la creación de mapas semánticos (Haspelmath 2003, Narrog 2010) en los que se plasma gráficamente la relación entre estas funciones que la morfología vincula por estar expresadas de la misma forma y queda representada su naturaleza de *continuum* (Crespo 1997). Si esta complejidad de notación en la que se incluyen las funciones semánticas, la sintaxis y la morfología se aplica a un corpus que permita hacer análisis diacrónico, los mapas semánticos resultantes pueden incluir vectores de direccionalidad, lo que ayuda a comprender el desarrollo de nuevas marcas morfológicas para las funciones semánticas, una vez que las marcas (morfológicas) originales han dejado de ser útiles por ser poco claras al ampliar su ámbito semántico de aplicación. En el capítulo VI, «La expresión morfológica de las funciones semánticas. Otro enfoque sobre las relaciones», Berta González se propone representar dichos fenómenos a través del análisis semántico comparado de las preposiciones *ab* del latín y *ἐκ* del griego. Esta comparación tiene como finalidad comprobar los puntos de contacto entre algunas funciones semánticas en estas dos lenguas, lo que refrenda estudios de tipología lingüística en esta misma línea y establece nuevos vínculos entre funciones semánticas que no aparecen recopilados en los repertorios tipológicos que recogen procesos de cambio semántico.

Finalmente, en el capítulo VII, «¿Adjunto o disjunto? Los sintagmas de *περί* + genitivo como expresión de la Referencia», Alberto Pardal propone un estudio sincrónico y diacrónico de las secuencias de *περί* + genitivo, que suelen aparecer en predicaciones con verbos de comunicación o de actividad mental (hablar / considerar acerca de algo), pero también, en ocasiones, es posible encontrar estas mismas expresiones en la periferia izquierda de la predicación, aparentemente estableciendo el tema de una unidad discursiva, esto es, desempeñando la función semántica Referencia. El objetivo de su estudio es dilucidar la naturaleza de estos constituyentes en relación con el marco predicativo de los verbos con los que habitualmente se construyen para determinar si son argumentales o no y explorar si estos sintagmas se están especializando mediante un proceso de pragmaticalización en el que estarían desarrollando funciones tematizadoras y/o topicalizadoras. En este segundo caso, es pertinente observar cuál es el proceso de evolu-

ción a partir de posibles contextos puente donde la frontera entre adjunto y disjunto resulta borrosa. La difícil determinación de la naturaleza sintáctico-semántica de estos elementos presenta, por último, retos para la anotación en un corpus digital, al tiempo que parte de la información que aportan (la pragmática) queda todavía fuera del alcance de la mayor parte de los métodos de anotación XML.

Como se puede apreciar, los aspectos que se abordan en este volumen pretenden dar una visión lo más amplia posible de las principales aportaciones que el proyecto COMREGLA quiere hacer en los estudios de lingüística en las lenguas clásicas. Este sistema permite un análisis morfológico, sintáctico y también semántico de los componentes que integran el texto, con una perspectiva novedosa que permite abrir un nuevo camino en el desarrollo de la lingüística de corpus para las lenguas clásicas.