# CORPUS NOTARIAL Y SINTÁCTICO DEL ASTURIANO MEDIEVAL (CONSAM-XIII)

# SYNTACTIC NOTARIAL CORPUS OF MEDIEVAL ASTURIAN (CONSAM-XIII)

**ROSABEL SAN SEGUNDO-CACHERO**
UNIVERSIDAD DE ZARAGOZA

ARTÍCULO RECIBIDO: 30-09-2017 | ARTÍCULO ACEPTADO: 01-12-2017

RESUMEN:
El *Corpus Notarial y Sintáctico del Asturiano Medieval (CoNSAM-XIII)* es un repositorio sintáctico integrado por 50 documentos notariales del siglo XIII procedentes del Archivo Catedralicio de Oviedo y del Archivo Municipal de Avilés en el que es posible buscar construcciones específicas y controlar múltiples variables al mismo tiempo mediante la combinación de criterios lineales y estructurales. Para su creación se han utilizado herramientas informáticas de software libre y un sistema de anotación sintáctica (Magro, Galves y Carrilho, 2016) basado en el estándar Penn-Helsinki (Kroch y Taylor, 2000) que permite la consulta mediante el motor de búsqueda *CorpusSearch* (Randall, 2005) y facilita la comparación con otras lenguas anotadas con el mismo sistema.

ABSTRACT:
The *Syntactic Notarial Corpus of Medieval Asturian (CoNSAM-XIII)* is a syntactic repository made up of 50 notarial deeds from the 13th century, kept in the Oviedo Cathedral Archive and the Avilés Municipal Archive. It enables us to search specific constructions and to select multiple variables simultaneously by combining linear and structural criteria. In order to create it free software tools and a syntactic annotation system (Magro, Galves y Carrilho, 2016) based on the standard Penn-Helsinki (Kroch y Taylor, 2000) have been used. This annotation system allows for queries and comparative studies among

languages annotated according to the same standard by using the searcher *CorpusSearch* (Randall, 2005).

PALABRAS CLAVE:
Sintaxis diacrónica, lingüística de corpus, anotación sintáctica, asturiano medieval, documentos notariales

KEYWORDS:
Diachronic Syntax, Corpus Linguistics, parsing, Medieval Asturian, notarial documents

---

**Rosabel San Segundo-Cachero.** Licenciada y doctora en Filología Hispánica por la Universidad de Oviedo. Su línea de investigación tiene como eje principal la variación sintáctica (sincrónica y diacrónica) y el cambio lingüístico en español y otras lenguas románicas peninsulares. Ha trabajado en el Consejo Superior de Investigaciones Científicas y en la Universidad de Lisboa y desde 2016 es profesora ayudante doctora en la Universidad de Zaragoza.

La creación del corpus es parte del proyecto *Sintaxis diacrónica del asturleonés (s. XIII): estructura oracional y orden de constituyentes* (ACA14-11), desarrollado en el Centro de Linguistica da Universidade de Lisboa y financiado por la Comisión Europea y el Gobierno del Principado de Asturias a través del programa de ayudas postdoctorales Marie-Curie-Clarín-COFUND (2014-2016).
Quiero expresar mi gratitud al Centro de Lingüística da Universidade de Lisboa por poner a mi disposición todos los medios necesarios para el desarrollo del proyecto y especialmente a Ana Maria Martins y a Sandra Pereira. Agradezco también a Xosé Lluis García Arias su ayuda con la interpretación de algunas palabras de los textos. No obstante, solo yo soy responsable de cualquier error o incoherencia que pueda haber en el análisis.

## 1. Introducción

La investigación lingüística se ha beneficiado en las últimas décadas de los avances de las nuevas tecnologías y de las herramientas informáticas que se han ido desarrollando. Así, en el caso del español ha sido posible la creación de grandes corpus de diverso tipo y finalidad como el *CORDE* (RAE), el *CREA* (RAE), el *Corpus del Español* (Davies, 2001-2016), la *Biblia Medieval* (Enrique-Arias y Pueyo Mena) o el *CODEA+2015* (GITHE), que son en la actualidad obras de referencia para cualquier lingüista. No obstante, la obtención de datos para el estudio de la sintaxis sigue siendo tarea ardua, porque, aunque algunos corpus permiten consultar ciertas combinaciones de unidades, no es posible buscar estructuras sintácticas mediante criterios de dependencia estructural.

En este sentido, la lingüística portuguesa ha avanzado notablemente, ya que cuenta con un corpus histórico desarrollado en Brasil, *Corpus Histórico do Português Tycho-Brahe (Tycho-Brahe)* (Galves y Britto, 2010) y otro, en Lisboa por el equipo *WOChWEL* (Martins, 2015), un corpus dialectal sincrónico del área portuguesa *Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)* (Martins, 2010), y un corpus epistolar luso-hispánico, *Post Scriptum (PS)* (CLUL, 2014) cuyos sistemas de etiquetado y anotación morfosintácticos son una adaptación del estándar Penn-Helsinki (Kroch y Taylor, 2000, Kroch *et. al.,* 2004, 2016, Santorini, 2016), lo que facilita la consulta de los corpus y favorece los estudios comparativos con lenguas que empleen en mismo sistema de anotación. Por ello, cuando comencé a estudiar la sintaxis del romance asturiano medieval y me percaté de la ausencia

de recursos informatizados, decidí crear un pequeño corpus sintáctico que resultase representativo del periodo estudiado y adoptar la metodología de los citados corpus portugueses.

En esas circunstancias y con ese objetivo surge el *Corpus Notarial y Sintáctico del Asturiano Medieval (CoNSAM-XIII),* que contiene 50 documentos notariales del siglo XIII, procedentes del Archivo Catedralicio de Oviedo (ACO) y del Archivo Municipal de Avilés (AMA). En las siguientes páginas describiré cómo se han procesado los textos y qué tipo de información morfológica y sintáctica se codifica en el corpus, así como la clase de consultas que se pueden realizar y las ventajas de usar un recurso de este tipo para estudiar la sintaxis.

## 2. ¿Por qué un corpus notarial?

La importante labor legislativa, cultural y filológica llevada a cabo durante el reinado de Fernando III y Alfonso X convierte al romance castellano y a la escritura toledana en el modelo lingüístico de la época, que se impone en los documentos oficiales de la cancillería castellano-leonesa. Sin embargo, en Asturias, la zona más periférica del reino, el romance asturleonés se mantiene en los documentos de carácter privado y local, escritos por notarios, cuya formación se supone inferior a la de los altos funcionarios de la cancillería. Por lo tanto, los únicos textos disponibles actualmente para estudiar el romance medieval asturiano son los documentos notariales escritos entre la segunda mitad del siglo XIII (antes solo hay fragmentos romances en textos latinos) y finales del siglo XIV, pues en el siglo XV concluye el proceso de castellanización y el asturleonés desaparece del registro escrito hasta siglos después (Morala, 2004 y García Arias, 2013).

Lejos de lo que pueda parecer por sus características textuales (Marín Martínez, 1991, Díez de Revenga, 1994, García Valle, 2004), el documento notarial constituye una importante fuente de información lingüística para el estudio de la sintaxis diacrónica y de la dialectología histórica, no solo por su abundancia y similitud en todo el territorio peninsular y en los distintos dominios lingüísticos, sino también porque las constricciones del patrón textual limitan las posibilidades interpretativas de las estructuras sintácticas y de la disposición de constituyentes, al mismo tiempo que su naturaleza performativa (Marín Martínez, 1991, Bono Huerta, 1992) lo hace permeable a rasgos diatópicos, diafásicos y diastráticos que no están presentes en otros textos (Menéndez Pidal, 1926).

## 3. Descripción y composición del corpus

La creación del corpus *CoNSAM-XIII* es, como he apuntado antes, consecuencia de la necesidad contar con una muestra representativa de las características sintácticas del asturleonés del siglo XIII, por lo que se han descartado textos de la primera mitad del siglo, cuya sintaxis es fundamentalmente latina, y los que no fueron escritos por notarios asturianos.

El *CoNSAM-XIII* está integrado por 50 documentos notariales originales,[1] 36 procedentes del ACO (testamentos, cartas de

[1] Una primera versión morfológica y sintáctica de los textos está disponible en el Respositorio Institucional del Principado de Asturias <http://ria.asturias.es/RIA/index.jsp> *(POS-tagged documents from the Oviedo Cathedral Archive, POS-tagged documents from the Avilés Municipal Archive, Parsed documents from the Oviedo Cathedral Archive, Parsed*

donación y venta) y 14, del AMA (cartas de vecindad, pleitos vecinales), todos ellos editados por Menéndez Gómez (2008) y publicados por la Academia de la Llingua Asturiana. Lo que ofrece este corpus no es, pues, una nueva edición de los textos, sino un repositorio de estructuras sintácticas, para cuya creación ha sido necesario modificar y simplificar el formato de la edición paleográfica utilizada en aras de obtener un texto compatible con las herramientas utilizadas para su procesamiento morfológico y sintáctico, como explicaré en el siguiente apartado. Por lo tanto, el lector no debe esperar unos resultados como los que ofrecen el *CODEA+2015* o *PS*, en los que es posible ver el original y distintos tipos de ediciones y presentaciones.

Lo que ofrece el *CoNSAM-XIII* es un conjunto de ficheros constituido por dos versiones de cada documento que corresponden a sendas fases del procesamiento computacional:

a. Versión morfológica: los textos únicamente llevan etiquetas morfológicas del sistema estandarizado POS *(part of speech)* pero no han sido sometidos al analizador sintáctico o *parser*. Convertida en TXT, esta versión puede ser utilizada con otros analizadores o puede convertirse a otro formato de etiquetado morfológico.

b. Versión sintáctica: los ficheros contienen el análisis sintáctico de los documentos, oración por oración, representado mediante estructuras parentéticas. Convertidos en ficheros PSD, son compatibles con el buscador *CorpusSearch* (Randall, 2005).

---

*documents from the Avilés Municipal Archive).* Aquí se presenta una versión revisada, corregida e individualizada de cada uno de los textos.

Los ficheros se agrupan en dos carpetas: una para los que contienen la versión morfológica *(CoNSAMXIII_POS)* y otra para los que contienen la versión sintáctica *(CoNSAMXIII_SYP)*. Para su correcta identificación, los ficheros llevan por nombre la referencia oficial de los manuscritos, a la que se añaden las codas *_POS* y *_SYP (syntactic parsing)*, respectivamente. El listado completo puede consultarse en el anexo 1, donde cada texto aparece con la referencia identificativa oficial y la referencia simplificada (conforme a los requisitos de las herramientas computacionales) que se utiliza dentro de los textos ya procesados para identificar y numerar todos los párrafos.

En total se han etiquetado y analizado morfosintácticamente 23.529 palabras: 17.130 corresponden a los 34 textos del ACO y 6.399, a los 16 textos del AMA. Dada su importancia histórica y lingüística, los documentos notariales que integran el *CoNSAM-XIII* se han analizado en su totalidad, sin omitir partes formulares ni fragmentos en latín, ya que los formulismos también sufren cambios a lo largo del tiempo y se observan diferencias entre notarios, lo que puede resultar de interés para distintos tipos de investigaciones.

## 4. Procesamiento de los textos

El procesamiento morfosintáctico se inicia a partir de un texto sin formato y se realiza en tres etapas en las que se utilizan sendas herramientas informáticas:

1. etiquetado morfológico POS *(part of speech)* con *eDictor* (Faria, Kepler & Paixão de Sousa, 2010);
2. análisis sintáctico automático *(parsing)* con el *parser* de Bikel (2004a, b), un analizador sintáctico de base

estadística que segmenta automáticamente los textos ya etiquetados por *eDictor* y establece una delimitación de los constituyentes;

3. revisión y corrección manual mediante la herramienta *CorpusDraw,* un complemento de *CorpusSearch* (Randall, 2005), que permite la visualización del texto en estructuras arbóreas editables.

Estas herramientas fueron diseñadas para los corpus diacrónicos del inglés, (Kroch y Taylor, 2000, Kroch *et. al.,* 2004, 2016), en la Universidad de Pensilvania y posteriormente se adaptaron al portugués para elaborar los corpus *Tycho-Brahe, CORDIAL-SIN, WOChWEL* y *PS*. Las etiquetas que se utilizan, tanto las morfológicas como las sintácticas, constituyen un sistema estandarizado y pueden emplearse de forma generalizada en las lenguas iberorrománicas para delimitar los constituyentes y los niveles estructurales, sin que ello implique renuncia alguna a reflejar las particularidades sintácticas de cada lengua, pues es tarea del investigador o del equipo de investigación establecer unos criterios para representar las estructuras sintácticas y para determinar el etiquetado que se va a emplear. Explicaré a continuación en qué consiste cada etapa del procesamiento y qué pautas se han seguido para la asignación de etiquetas.

### 4.1. El texto sin formato: la presentación crítica

El paso previo al etiquetado morfológico es la conversión de la edición paleográfica de los documentos (Menéndez Gómez, 2008) en un texto legible por el etiquetador morfosintáctico de *eDictor*. *Vid*. (1) y (2). Para ello es necesario eliminar todas las marcas tipográficas (cursiva, superíndice, subíndice, etc.), regular

la puntuación y el uso de las mayúsculas para tratar de reflejar la sintaxis del original y facilitar la segmentación de constituyentes. El resultado es algo similar a lo que se denomina "presentación crítica" (AA.VV., 2013): un texto que se mantiene fiel al original, pero que ha sido regularizado en mayor o menor medida de acuerdo a una determinada finalidad. En el caso del *CoNSAM-XIII* se ha querido alterar lo menos posible la edición paleográfica y, por ello, se han aplicado los siguientes criterios:

a) no se alteran las grafías ni la acentuación de la edición paleográfica utilizada;

b) se eliminan las marcas de edición, a excepción de los tres puntos entre corchetes […] que señala las partes del texto perdidas o no legibles;

c) se regulariza el uso de las mayúsculas conforme a los criterios actuales;

d) el signo tironiano se desarrolla siempre como *et* para evitar decantarse por una forma concreta de la conjunción copulativa, ya que en los textos puede aparecer de distintas formas *(et, e, hie)*;

e) se intenta reflejar en la medida de lo posible la sintaxis de la época mediante el sistema de puntuación actual; pero

f) se mantiene la unión y separación de palabras de la edición paleográfica y se pospone hasta la fase de análisis sintáctico la delimitación de las unidades lingüísticas conforme a los criterios gramaticales actuales, pero sin perder información con respecto a la unión que pudiera existir en el original.

(1) Fragmento de la edición paleográfica (Menéndez Gómez ,2008: 49):

[…] mandamos p*er* manda τ so pena de la /[12] fiadoria q*ue* se contie*n* en el co*m*promisso q*ue* sont dos mill m*a*r*a*v*edis* dela moneda noua./ q*ue*l Concello$_{w}$ de abilles entregue*n* luego alos vezinos$_{x}$ de Ouiedo los pa*n*nos q*ue*llos p*r*indaron (AMA, 23, 1289)

(2) Fragmento de la presentación crítica:

[…] mandamos per manda et so pena de la fiadoria que se contien en el compromisso, que sont dos mill maravedis dela moneda noua, quel Concello de Abilles entreguen luego alos vezinos de Ouiedo los pannos quellos prindaron (AMA, 23, 1289)

### 4.2. El etiquetado morfosintáctico

Una vez preparado, el texto sin formato es procesado por el etiquetador de *eDictor,* que asigna a cada palabra una etiqueta morfosintáctica conforme al sistema estandarizado POS utilizado para el portugués (Britto *et al*. 2016, Magro y Morgado, 2008, Martins, 2015).

Como se puede ver en las tablas del anexo 2, las etiquetas POS indican la clase, léxica o funcional, a la que pertenece cada palabra,[2] así como los rasgos flexivos y/o semánticos asociados a cada categoría, que se expresan mediante subetiquetas. Por lo que respecta a los verbos, se establece una distinción entre los verbos léxicos /VB y los que pueden utilizarse también como auxiliares en función del contexto (*ser*/SR, *estar*/ET, *haber*/HV, *tener*/TR) y solo se reflejan cuatro tiempos verbales, cuyas etiquetas pueden

[2] En casos ambiguos o dudosos se opta por aplicar la etiqueta que parece más adecuada según el contexto sintáctico o según el patrón textual.

combinarse con la etiqueta de subjuntivo /-S. Obsérvese que el etiquetado no refleja los rasgos considerados no marcados, como el género masculino, el número singular o el modo indicativo.

La configuración del etiquetador *eDictor* permite asignar a un segmento lingüístico integrado por distintas unidades una etiqueta compleja, compuesta por tantas etiquetas como unidades contenga el segmento lingüístico en cuestión, separadas por el signo +. Sin embargo, dado que el *parser* no puede leer estas etiquetas complejas, como paso previo al análisis sintáctico, es necesario separar las palabras que, respetando la edición paleográfica, se han mantenido juntas. Para separarlas se adoptan criterios lexicológicos vigentes en la época del texto y se les asignan individualmente las etiquetas, dejando constancia de la unión gráfica de las palabras mediante la inserción del símbolo @ en el punto de unión.

(3) Fragmento versión morsintáctica (AMA,23,1289_POS)

mandamos/VB-P per/P manda/N et/CONJ so/P pena/N de/P la/D-F fiadoria/N que/WPRO se/SE contien/VB-P en/P el/D compromisso/N ,/, que/WPRO sont/SR-P dos/NUM mill/NUM maravedis/N-P **dela/P+D-F** moneda/N noua/ADJ-F ,/, **quel/C+D** Concello/NPR de/P Abilles/NPR entreguen/VB-SP luego/ADV **alos/P+D-P** vezinos/N-P de/P Ouiedo/NPR los/D-P pannos/N-P **quellos/WPRO+CL** prindaron/VB-D (AMA1289,1.10)

(4) Fragmento convertido en input del *parser*:

mandamos/VB-P per/P manda/N et/CONJ so/P pena/N de/P la/D-F fiadoria/N que/WPRO se/SE contien/VB-P en/P el/D compromisso/N ,/, que/WPRO sont/SR-P dos/NUM mill/NUM

maravedis/N-P **de@/P @la/D-F** moneda/N noua/ADJ-F ,/, **que@/C @l/D** Concello/NPR de/P Abilles/NPR entreguen/VB-SP luego/ADV **a@/P @los/D-P** vezinos/N-P de/P Ouiedo/NPR los/D-P pannos/N-P **que@/WPRO @llos/CL** prindaron/VB-D (AMA1289,1.10)

### 4.3. El análisis sintáctico

Una vez adaptado el *output* de *eDictor* a las condiciones de legibilidad del *parser* (Bikel, 2004a), se lleva a cabo el análisis sintáctico de forma automática. El *parser* es un analizador de base estadística que lee las etiquetas morfosintácticas y establece los límites entre constituyentes, atribuyéndoles una estructura y asignándoles unas etiquetas sintácticas basadas en un modelo formalista que refleja la proyección de categorías léxicas y funcionales y especifica su función sintáctica.

Como el analizador fue concebido para que sus resultados pudieran ser aceptados por la mayoría de los modelos teóricos, no hay ramificación binaria de las estructuras (como en los modelos formalistas más extendidos) y la categoría verbal no proyecta, de manera que el verbo y todos sus argumentos y satélites están dominados por el nudo IP *(Inflectional Phrase)*. También se utilizan categorías vacías, cuyo uso combinado con la coindexación permite reflejar el movimiento de constituyentes, sin que ello implique la adopción de un modelo teórico que asuma la existencia de tales categorías. Estas etiquetas se utilizan como recurso o estrategia para localizar elementos desplazados y construcciones que de otro modo no podríamos encontrar, como la subida de clíticos, movimiento de sintagmas interrogativos, etc. Recordemos que el objetivo de un corpus sintáctico no es otro que proporcionar

material para investigaciones lingüísticas, por lo que no se pretende condicionar la interpretación de los datos, sino facilitar su localización. En el anexo 3 pueden consultarse las etiquetas utilizadas en el *CoNSAM-XIII,* un subconjunto de la adaptación portuguesa del sistema de anotación sintáctica *Penn-Helsinki* realizada por equipos brasileños y portugueses (Galves 2008, Carrilho y Magro, 2011,) y unificada en Magro, Galves y Carrillo (2016).

El análisis sintáctico se representa en estructuras parentéticas (Figura 1) que pueden visualizarse como árboles sintácticos (Figura 2) si los ficheros PSD se abren con el editor *CorpusDraw* (Randall, 2005), una herramienta que permite la revisión y la corrección manual de las estructuras generadas automáticamente por el *parser*.

```
(CP-THT (CP-THT (C que@)
                (IP-SUB (NP-SBJ (D @l)
                                (NPR Concello)
                                (PP (P de)
                                    (NP (NPR Abilles))))
                        (VB-SP entreguen)
                        (ADVP (ADV luego))
                        (PP (P a@)
                            (NP (D-P @los)
                                (N-P vezinos)
                                (PP (P de)
                                    (NP (NPR Ouiedo)))))
                        (NP-ACC (D-P los)
                                (N-P pannos)
                                (CP-REL (WNP-7 (WPRO que@))
                                        (IP-SUB (NP-ACC *T*-7)
                                                (NP-SBJ *pro*)
                                                (NP-DAT (CL @llos))
                                                (VB-D prindaron))))))))
```

Figura 1. Análisis sintáctico: estructura parentética

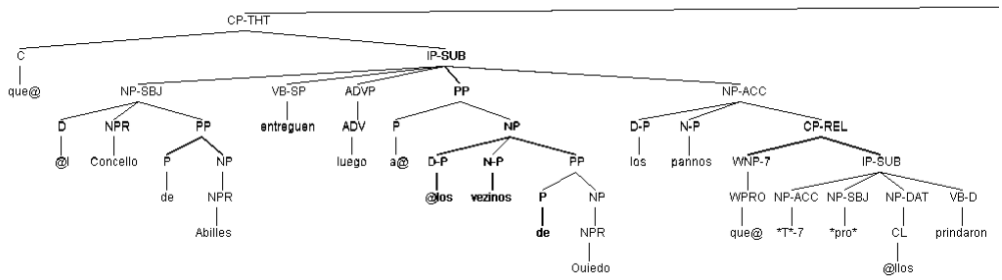


Figura 2. Análisis sintáctico: estructura arbórea

### 4.4. El análisis sintáctico en el CoNSAM-XIII: algunas precisiones

Aunque en el *CoNSAM-XIII* se utiliza el sistema de anotación desarrollado para el portugués, hay algunas divergencias en cuanto la representación de determinadas estructuras que conviene tener en cuenta a la hora de realizar búsquedas con *CorpusSearch* para estudios comparativos.

Las oraciones completivas no interrogativas seleccionadas por un verbo o por una preposición se etiquetan siempre como CP-THT; por el contrario, en los corpus portugueses se utiliza CP-ADV cuando la oración subordinada está dominada por una preposición y todo el segmento desempeña una función no argumental. Así, *para que* sería (PP (P) (CP-THT)) en el corpus asturiano y (PP (P) (CP-ADV)) en los corpus portugueses. En el caso de las oraciones con valor causal, si la preposición y el complementizador están escritos como dos unidades independientes, *por*/P *que*/C, se analizan igual que *para que* (PP (P) (CP-THT)); mientras que, si ambas constituyen una única unidad (*porque*/C), entonces la subordinada es (CP-ADV).

Entre las oraciones relativas, las libres y semilibres se anotan de forma diferente a las relativas dependientes. Estas últimas llevan la etiqueta CP-REL y están dominadas por la proyección máxima del núcleo al que complementan (NP-ACC (N) (CP-REL)). Las otras dos, en cambio, se etiquetan como CP-FRL y no son hermanas de ningún núcleo. Si se trata de una relativa semilibre (*el que, las que,* etc.) o de una relativa libre encabezada por el pronombre *quien,* el nudo que domina a CP-FRL es NP, acompañado de la subetiqueta que le corresponda según su función sintáctica. Si las

relativas están encabezadas por un adverbio relativo *(donde, cuando, como, mientras, según),* en el *CoNSAM-XIII* se consideran relativas libres que desempeñan funciones adverbiales (Brucart, 1999), por lo que el nudo que las domina es ADVP. En cambio, en los corpus portugueses solo se procede de este modo cuando el adverbio relativo es el locativo *onde,* pues las subordinadas temporales se etiquetan como CP-ADV y las modales introducidas por *como,* aunque se etiquetan como CP-ADV, están precedidas de un constituyente adverbial vacío. En el *CoNSAM-XIII,* pues, se ha adoptado una representación única para todos estos casos que parece reflejar de forma adecuada las características formales de estas estructuras con independencia de su contenido semántico.

## 5. Búsquedas automáticas con *CorpusSearch*

La consulta automática del corpus está supeditada a la conversión de los ficheros en archivos PSD que sean legibles por el motor de búsqueda *CorpusSearch* (Randall, 2005).[3] El modo de hacer las consultas o *queries* consiste en crear ficheros de texto y combinar en ellos operadores lógicos con comandos de búsqueda (basados en criterios de linealidad y jerarquía estructural) y etiquetas morfosintácticas. En (5) y (6) se presentan algunos ejemplos donde aparece la *query* con su glosa correspondiente y el resultado obtenido: en (5) se buscan estructuras en las que el OD es una completiva postverbal, y en (6) oraciones con ascenso de clíticos. El producto de la búsqueda es un nuevo fichero de texto en el que aparecen las estructuras solicitadas con su correspondiente análisis y el recuento total de los resultados. Como explico en San-

[3] Los textos en formato PSD pueden solicitarse a la autora.

Segundo-Cachero (2017) con un caso práctico, es imprescindible conocer bien el sistema de anotación sintáctica para poder realizar consultas precisas y evitar resultados no deseados. Puede consultarse una guía detallada para la realización de *queries* en http://corpussearch.sourceforge.net/ .

(5) Consulta 1: V+completiva

```
node: $ROOT
query: ({1}IP-SUB idoms {2}VB*|HV*|TR*)
AND ({1}IP-SUB idoms {3}CP-THT|CP-QUE)
AND ({2}VB*|HV*|TR* Precedes {3}CP-THT|CP-QUE)
```

Figura 3. *Query* 1

Glosa: oración subordinada con verbo no copulativo que domina a una oración completiva o interrogativa que aparece en posición postverbal.

Resultado: "Et faggo este otro que mando que sea firme et uala". (ACO1289,3,.9)

```
( (IP-MAT (CONJ Et)
          (NP-SBJ *pro*)
          (VB-P ffago)
          (NP-ACC (D este)
                  (OUTRO otro)
                  (CP-REL (WNP-1-2 (WPRO que))
                          (IP-SUB (NP-SBJ *pro*)
                                  (VB-P mando)
                                  (CP-THT (C que)
                                          (IP-SUB (IP-SUB (NP-SBJ *T*-1)
                                                          (SR-SP sea)
                                                          (ADJP (ADJ-G firme)))
                                                  (CONJP (CONJ et)
                                                         (IP-SUB (NP-SBJ *T*-2)
                                                                 (VB-SP uala))))))))
          (. .))
```

Figura 4. Resultado de *query* 1

(6) Consulta 2: ascenso de clíticos

```
node: $ROOT
query: ({1}NP* sameIndex {2}\*-*)
AND ({1}NP* Dominates {4}CL)
AND ({1}NP* HasSister {3}VB-*)
AND ({1}NP* Precedes {3}VB-*)
```

Figura 5. *Query* 2

Glosa: en cualquier oración, un sintagma nominal constituido por un clítico está coindexado con una huella de clítico y tiene como hermano un verbo finito al que precede.

Resultado: "por otros que los podan scomungar" (ACO1281,1,9)

```
(PP (P por)
    (NP (D-P los)
        (OUTRO-P otros)
        (CP-REL (WNP-2 (WPRO que))
                (IP-SUB (NP-SBJ *T*-2)
                        (NP-3 (CL los))
                        (VB-SP podan)
                        (IP-INF (NP-ACC *-3)
                                (VB scomungar)))))
```

Figura 6. Resultado de *query* 2

## 6. Conclusiones

Aunque de reducido volumen, el corpus *CoNSAM-XIII* ofrece un repositorio de estructuras sintácticas representativas del romance asturleonés del siglo XIII procedentes de documentos notariales asturianos originales. No es, por supuesto, un trabajo concluido, sino un primer paso en la creación de una base de datos sintáctica más amplia. La descripción de la metodología y de las herramientas de acceso libre y gratuito utilizadas pretende ser un

acicate para animar a quienes trabajan en el campo de la sintaxis a crear corpus sintácticos de diversas épocas, lenguas y tipologías textuales que faciliten el estudio comparado de la sintaxis dialectal, tanto diacrónica como sincrónica, de los romances peninsulares.

## 7. Referencias

Bikel, Dan (2004a). *dbparser*. Ubuntu 8.04 LTS. 32-bits. <http://www.tycho.iel.unicamp.br/~tycho/apps/dbparser-files/>.

Bikel, Dan (2004b). *On the parameter space of generative lexicalized statistical parsing models*. Tesis doctoral University of Pennsylvania. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.2734&rep=rep1&type=pdf>.

Bono Huerta, José (1992). "Conceptos fundamentales de la diplomática notarial". *Historia. Instituciones. Documentos* 19: pp. 73-88.

Britto, Helena, Maria Clara Paixão de Sousa, Shirley Guedes y Charlotte Galves (2016). *The Tycho Brahe Corpus Annotation System. Morphological Tags (POS and Inflectional)*. Universidade Estadual de Campinas (UNICAM). <http://www.tycho.iel.unicamp.br/corpus/manual/pos2016.html>. (26-11-2016).

Carrilho, Ernestina y Catarina Magro (2011). *CORDIAL-SIN Syntactic Annotation System Manual*. Centro de Linguistica, Universidade de Lisboa. <http://www.clul.ul.pt/cordial-sam/>. (24-7-2017).

Centro de Linguistica da Universidade de Lisboa (CLUL) (2014). *P. S. Post Scriptum. Arquivo digital de escrita quotidiana em Portugal e Espanha na época moderna*. <http://ps.clul.ul.pt>. (6-6-2017).

Davies, Mark. (2001-2016). *Corpus del Español*. <https://www.corpusdelespanol.org/x.asp>. (15-9-2017).

Díez de Revenga, Pilar (1985). "Análisis de las lexías complejas en documentos medievales murcianos". *Estudios Lingüísticos de la Universidad de Alicante* 3: pp. 193-208.

Enrique-Arias, Andrés, y F. Javier Pueyo Mena. *Biblia Medieval* [en línea]. <http://www.bibliamedieval.es>. (15-9-2017).

Faria, Pablo, Fábio Kepler y Maria Clara Paixão de Sousa (2010). *eDictor* 1.0 beta10. <https://humanidadesdigitais.org/edictor/>. (7-8-2016).

Galves, Charlotte (2008). *Tycho Brahe Parsed Corpus of Historical Portuguese. Syntactic Annotation System*. Universidade Estadual de Campinas (UNICAM). <http://www.tycho.iel.unicamp.br/corpus/manual/syn-frm.html>. (18-5-2017).

Galves, Charlotte y Pablo Faria (2010). *The Tycho Brahe Corpus of Historical Portuguese*. Universidade Estadual de Campinas (UNICAM). <http://www.tycho.iel.unicamp.br/~tycho/>. (20-6-2017).

García Arias, Xosé Lluis (2013). "Conciencia llingüística y testos asturianos medievales". *Lletres Asturianes. Boletín de l'Academia de la Llingua Asturiana* 108: pp. 87-106.

García Valle, Adela (2004). "Las fórmulas jurídicas medievales. Un acercamiento preliminar desde la documentación notarial

de Navarra". *Anuario de historia del derecho español* 74: pp. 613-640.

GITHE (Grupo de Investigación Textos para la Historia del Español). *CODEA+ 2015 (Corpus de Documentos Españoles Anteriores a 1800)* [en línea]. <http://corpuscodea.es>. (15-9-2017).

Kroch, Anthony, Beatrice Santorini y Lauren Delfs (2004). *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. <http://www.ling.upenn.edu/ppche-release-2016/PPCEME-RELEASE-3>. (20-6-2017).

Kroch, Anthony, Beatrice Santorini y Ariel Diertani (2016). *The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE2)*. <http://www.ling.upenn.edu/ppche-release-2016/PPCMBE2-RELEASE-1>.(20-6-2017).

Kroch, Anthony y Ann Taylor (2000). *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. <http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4>. (20-6-2017).

Magro, Catarina, Charlotte Galves y Ernestina Carrilho (2016). *Portuguese Syntactic Annotation Manual*. Lisboa/ Campinas: Centro de Linguística da Universidade de Lisboa/ Instituto de Estudos da Linguagem da Universidade de Campinas. Ms.

Magro, Catarina y Cristina Morgado (2008). *CORDIAL-SIN POS Annotation Manual*. Centro de Linguística, Universidade de Lisboa. <http://www.clul.ul.pt/english/sectores/variacao/cordialsin/pos_annotation_manual.pdf>. (2-9-2017).

Martins, Ana Maria (2015). *Word Order and Word Order Change in Western European Languages*. Centro de Linguística,

Universidade de Lisboa. <http://alfclul.clul.ul.pt/wochwel/oldtexts.html>. (2-9-2017).

Martins, Ana Maria (coord.) (2000-2010). *CORDIAL-SIN: Corpus Dialectal para o Estudo da Sintaxe*. Centro de Linguística, Universidade de Lisboa. <http://www.clul.ul.pt/en/resources/411-cordial-corpus>. (6-5-2017).

Menéndez Gómez, Jesús (ed.) (2008). D*ocumentos orixinales del dominiu llingüísticu ástur I. (1244-1299)*. Uviéu: Academia de la Llingua Asturiana.

Menéndez Pidal, Ramón (1926). *Orígenes del español. Estado lingüístico de la Península Ibérica hasta el siglo XI*. Madrid: Espasa, ed. 1950.

Morala, José Ramón (2004). "Del leonés al castellano". Coord. Rafael Cano Aguilar. *Historia de la Lengua Española*. Barcelona: Ariel. pp. 555-570.

Randall, Beth (2005). *CorpusSearch 2*. Windows-64bits. Universiy of Pennsylvania. <http://corpussearch.sourceforge.net/>. (24-9-2016).

Real Academia Española. Banco de Datos *(CORDE)* [en línea]. *Corpus Diacrónico del Español*. <http://www.rae.es>. (15-9-2017).

Real Academia Española. Banco de Datos *(CREA)* [en línea]. *Corpus de Referencia del Español Actual*. <http://www.rae.es>. (15-9-2017).

San Segundo-Cachero, Rosabel. (2017). "La anotación sintáctica de textos medievales: un recurso fundamental para el estudio del orden de constituyentes". Ed. Silvia Gumiel.

*Investigaciones en Lingüística: Vol. III: Sintaxis*. Alcalá de Henares: Universidad de Alcalá de Henares.

Santorini, Beatrice. (2016). *Annotation Manual for the Penn Historical Corpora and the York-Helsinki Corpus of Early English Correspondence*. University of Pennsylvania. <http://www.ling.upenn.edu/ppche/ppche-release-2016/annotation/index.html> (17-8-2016)

Varios Autores (2013). "Criterios de edición de documentos hispánicos (Orígenes-Siglo XIX) de la Red Internacional CHARTA". *Corpus hispánico y americano en la red: textos antiguos*. <http://files.redcharta1.webnode.es/200000023-de670df5d6/Criterios%20CHARTA%2011abr2013.pdf>. (03-04-2016).

## 8. Anexo 1. Textos que integran el *CoNSAM-XIII*

Documentos descargables desde: http://revistacaracteres.net/wp-content/uploads/2018/05/CoNSAM-XIII.zip

| Referencia identificativa | Referencia simplificada |
|---|---|
| ACO, A.7.3, 1244 | ACO1244,1 |
| ACO, A.7.6, 1247 | ACO1247,1 |
| ACO, A.7.7, 1247 | ACO1247,2 |
| ACO, A.7.8, 1247 | ACO1247,3 |
| ACO, A.7.9, 1249 | ACO1249,1 |
| ACO, A.7.11, 1254 | ACO1254,1 |
| ACO, A.7.12, 1254 | ACO1254,2 |
| ACO, A. 7.13, 1255 | ACO1255,1 |
| ACO, A.7.16, 1257 | ACO1257,1 |
| ACO, A.8.2, 1260 | ACO1260,1 |
| ACO, A.8.11, 1268 | ACO1268,1 |
| ACO, A.8.12, 1269 | ACO1296,1 |
| ACO, A.8.13, 1269 | ACO1269,2 |
| ACO, A.8.16, 1271 | ACO1271,1 |
| ACO, A.9.1, 1272 | ACO1272,1 |
| ACO, A.9.2, 1273 | ACO1273,1 |
| ACO, A.9.6, 1274 | ACO1274,1 |
| ACO, A.9.7, 1275 | ACO1275,1 |

| | |
|---|---|
| ACO, A.9.8, 1277 | ACO1277,1 |
| ACO, A.9.9, 1277 | ACO1277,2 |
| ACO, A.9.10, 1278 | ACO1278,1 |
| ACO, A.9.11, 1278 | ACO1278,2 |
| ACO, A.9.12, 1281 | ACO1281,1 |
| ACO, A.9.13, 1283 | ACO1283,1 |
| ACO, A.9.14, 1285 | ACO1284,1 |
| ACO, A.9.15, 1285 | ACO1285,2 |
| ACO, A.9.16, 1286 | ACO1286,1 |
| ACO, A.10.1, 1286 | ACO1286,2 |
| ACO, A.10.2, 1287 | ACO1287,1 |
| ACO, A.10.3, 1287 | ACO1287,2 |
| ACO, A.10.4, 1289 | ACO1289,1 |
| ACO, A.10.5, 1289 | ACO1289,2 |
| ACO, A.10.6, 1289 | ACO1289,3 |
| ACO, A.10.7, 1289 | ACO1289,4 |
| AMA, 3, 1266 | AMA1266,1 |
| AMA, 4, 1269 | AMA1269,1 |
| AMA, 6, 1280 | AMA1280,1 |
| AMA, 7, 1281 | AMA1281,1 |
| AMA, 8, 1281 | AMA1281,2 |
| AMA, 9, 1281 | AMA1281,3 |
| AMA, 10, 1281 | AMA1281,4 |
| AMA, 12, 1282 | AMA1282,1 |

| | |
|---|---|
| AMA, 13, 1283 | AMA1283,1 |
| AMA, 14, 1284 | AMA1284,1 |
| AMA, 16, 1286 | AMA1286,1 |
| AMA, 17, 1286 | AMA1286,2 |
| AMA, 19, 1287 | AMA1287,1 |
| AMA, 23, 1289 | AMA1289,1 |
| AMA, 24, 1289 | AMA1289,2 |
| AMA, 27, 1299 | AMA1299,1 |

# 9. Anexo 2. Etiquetas POS

| Categorías léxicas | | Categorías funcionales | |
|---|---|---|---|
| Verbo pleno | /VB | Conjunción | /CONJ |
| Auxiliar/ pleno | /SR, /ET, /HV, /TR | Complementante | /C |
| Nombre | /N | Pronombre | /PRO |
| Adjetivo | /ADJ | Determinante | /D |
| Adverbio | /ADV | Posesivo | /PRO$ |
| Preposición | /P | Cuantificador | /Q |
| | | Negación | /NEG |

Figura 7. Clases de palabras

| | | Indicativo | Subjuntivo | Imperativo |
|---|---|---|---|---|
| Formas finitas | Presente | /VB-P | /VB-SP | /VB-I |
| | Pasado | /VB-D | /VB-SD | |
| | Futuro y condicional | /VB-R | /VB-SR | |
| | Formas con *-ra* | /VB-RA | | |
| Formas no finitas | Infinitivo | /VB, /SR, /ET, /HV, /TR | | |
| | Gerundio | /VB-G | | |
| | Participio | /VB-PP | | |
| | Participio activo | /VB-AG | | |

Figura 8. Flexión verbal

| | (Masculino singular) | Femenino | Género invariable | Plural |
|---|---|---|---|---|
| Nombre | /N | /N-F | | /N(-F)-P |
| Adjetivo | /ADJ | /ADJ-F | /ADJ-G | /ADJ(-X)-P |
| Determinante y demostrativos variables | /D | /D-F | /D-G | /D(-X)-P |
| Determinante indefinido / numeral | /D-UM | /D-UM-F | | /D-UM(-F)-P |
| Posesivo | /PRO$ | /PRO$-F | | /PRO$(-F)-P |
| *Otro,-a,-s* | /OUTRO | /OUTRO-F | | /OUTRO(-F)-P |
| Pronombre | /PRO | /PRO | | /PRO |
| Cuantificador | /Q | /Q-F | /Q-G | /Q(-X)-P |
| Demostrativo invariable | /DEM | | | |
| Pronombre clítico | /CL | /CL | /CL | /CL |
| *se* | /SE | /SE | /SE | /SE |

Figura 9. Flexión nominal

| | (Masculino) | Femenino | Género invariable | Plural |
|---|---|---|---|---|
| Adjetivo comparativo | /ADJ-R | /ADJ-R-F | /ADJ-R-G | /ADJ-R-(-F)-P |
| Adjetivo superlativo | /ADJ-S | /ADJ-S-F | /ADJ-S-G | /ADJ-S(-F)-P |
| Adverbio comparativo | /ADV-R | | | |
| Adverbio superlativo | /ADV-S | | | |

Figura 10. Categorías con gradación

| | Masculino | Femenino | Plural |
|---|---|---|---|
| Negación | /NEG | | |
| *Sinón*[4] | /SENAO | | |
| Cuantificador | /Q-NEG | /Q-NEG-F | /Q-NEG(-F)-P |
| Adverbio | /ADV-NEG | | |
| Conjunción | /CONJ-NEG | | |
| Partículas focales negativas | /FP-NEG | | |

Figura 11. Categorías negativas

| | Masculino | Femenino | Género invariable | Plural |
|---|---|---|---|---|
| Pronombres | /WPRO | /WPRO-F | | /PRO(-F)-P |
| Adjetivos[5] | /WADJ | /WADJ-F | /WADJ-G | /WADJ(-X)-P |
| Determinante | /WD | /WD-F | /WD-G | /WD(-X)-P |
| Posesivo | /WPRO$ | /WPRO$-F | | /WPRO$-F-P |
| Adverbio | /WADV | | | |

Figura 12. Unidades relativas e interrogativas

[4] Se adopta esta etiqueta para el elemento que introduce una construcción exceptiva: *senão* en portugués y *sinón* en asturiano.

[5] Etiqueta añadida en el *CoNSAM-XIII* por ser necesaria para representar algunas estructuras.

| Otros constituyentes | | Puntuación | |
|---|---|---|---|
| Partículas focales | /FP | Punto | ./. |
| Números cardinales | /NUM | Dos puntos | :/. |
| Interjecciones | /INTJ | Punto y coma | ;/. |
| Palabras desconocidas | /FW | Exclamación | !/. |
| Texto omitido […] | /CODE | Interrogación | ?/. |
| | | Coma | ,/, |
| | | Comillas | “/QT ”/QT |
| | | Paréntesis | (/( )/) |
| | | Guion | -/( -/) |

Figura 13. Otras etiquetas

| **Este mismo texto en la web** |
|---|
| http://revistacaracteres.net/revista/vol7n1mayo2018/consam-xiii |