Corpus electrónicos históricos y usuarios, con atención especial al *CORHEN*[1]

Historical electronic corpora and users, with special attention to *CORHEN*

Rosa M. Espinosa Elorza
Universidad de Valladolid (Valladolid, España)
respelorza@yahoo.es
https://orcid.org/0000-0001-8878-6885

Andrzej Zieliński
Uniwersytet Jagielloński (Kraków, Polonia), Uniwersytet Szczeciński (Szczecin, Polonia)
andrzej.zielinski@uj.edu.pl, andrzej.zielinski@usz.edu.pl
https://orcid.org/0000-0001-8609-0761

Recibido el 17/10/2023, aceptado el 6/1/2024, publicado el 18/10/2024

Cómo citar este artículo
Espinosa Elorza, Rosa M., Andrzej Zieliński 2024. Corpus electrónicos históricos y usuarios, con atención especial al *CORHEN*. *Studia linguistica romanica* 2024.12, 58-85. https://doi.org/10.25364/19.2024.12.4.

Resumen
Como usuarios de corpus electrónicos, conscientes de lo complejo que resulta elaborarlos y prepararlos para su consulta, queremos centrarnos en unos puntos que, con la ayuda de especialistas – ingenieros, informáticos, filólogos y lingüistas –, servirán en un futuro para afinar las herramientas que mejoren su uso. Los corpus electrónicos facilitan el trabajo especialmente a los lexicógrafos y a los expertos en terminología; sin embargo, quienes se dedican a la morfología o a la sintaxis han de emplear mucho tiempo en discriminar los elementos que les interesan. Lo ideal sería contar con etiquetas textuales, gráfico-fonéticas, morfológicas, sintácticas, semánticas y pragmáticas; asimismo, sería muy útil incluir las referencias a las relaciones entre documentos y la consulta de facsímiles o de la obra original de la que es traducción el texto manejado en un determinado momento.

1 Esta investigación se ha desarrollado en el ámbito del proyecto *El castellano norteño en la Edad Media*, dirigido por María J. Torrens Álvarez (Instituto de lengua, literatura y antropología del Consejo superior de investigaciones científicas). También fue financiada por el *Priority research area Heritage* dentro del programa *Initiative of excellence-Research university* de la Universidad Jaguelónica de Cracovia.

Abstract
As users of electronic corpora, and aware of the complexities involved in their development and preparation for consultation, we aim to highlight several points that, with the help of specialists – engineers, computer scientists, philologists and linguists –, can contribute to refining the tools that will enhance their usability in the future. Electronic corpora greatly facilitate the work of lexicographers and terminology experts; however, those focused on morphology or syntax often need to invest significant time in identifying the elements relevant to their research. Ideally, corpora would feature textual, graphic-phonetic, morphological, syntactic, semantic and pragmatic tags. Additionally, it would be very useful to include references to the relationships between documents, along with access to facsimiles or to the original works from which the texts under consideration have been translated.

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

Índice

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

1 Introducción

[1] Desde 1961, año en el que se creó el primer corpus electrónico, el *Brown University standard corpus of present-day American English*, ideado para analizar la frecuencia y la distribución de las categorías lingüísticas, han ido apareciendo otros con herramientas que proporcionan datos empíricos cada vez más fidedignos. En vez de la ardua lectura horizontal de diferentes producciones escritas por parte del investigador, en quien se puede producir la famosa paradoja de observador, como advirtió Labov (1972: 171) – esto es, la influencia del observador en lo observado –, con un solo clic en determinado corpus se obtienen miles de elementos aislados e incluso expresiones más o menos complejas. De este modo, los corpus electrónicos permiten pasar del análisis de la *parole* – una determinada producción individual – a la realización de investigaciones sobre la *langue* – el sistema lingüístico en conjunto – (García-Miguel 2022; Tognini-Bonelli 2001; entre otros).

[2] En lo que respecta a los estudios diacrónicos del español, el avance tecnológico que se experimentó desde finales de los años noventa del siglo pasado corrió parejo a la creación de varios corpus históricos que, a pesar de sus imperfecciones, indudablemente contribuyen a dar una imagen cada vez más pulida no solamente del funcionamiento de unidades lingüísticas utilizadas en épocas pasadas, sino también de su evolución. De hecho, el estudio diacrónico depende fundamentalmente de los datos extraídos de los corpus históricos digitalizados; en palabras de McEnery, Xiao & Tono (2006: 46), «diachronic study is perhaps one of the few areas which can only be investigated using corpus data».

[3] Para nuestra lengua, el primer banco de datos electrónico fue el *ADMyTE* en formato CD-ROM, editado por Francisco Marcos Marín, Charles Faulhaber y Ángel Gómez Moreno en 1991, que incluía unos trescientos textos. Unos años más tarde, la Real Academia Española puso en marcha el banco de datos de libre acceso *CORDE*, que, hasta la fecha, contiene 300 millones de formas desde los textos más antiguos hasta 1974, distribuidos en diferentes géneros discursivos y espacios geográficos (Sánchez Sánchez & Domínguez Cintas 2007: 138). En 2001-2002 fue publicado, también en línea, el *Corpus del español* (*CE*) por Mark Davies, que, en su vertiente diacrónica, cuenta con unos 100 millones de palabras pertenecientes a textos de varios géneros discursivos y distintos ámbitos geográficos desde el siglo XIII hasta el XX.

[4] La creación de corpus de tamaña extensión implica una notable pérdida de acotaciones, razón por la cual los bancos de datos electrónicos posteriores – no agotamos las citas – optaron por paliar esta carencia y centrarse:

a) En lo diatópico, como el *CORHEN*, con 253 documentos notariales de carácter particular del norte de Castilla – provincias actuales de Burgos, Santander y Palencia – desde el año 922 hasta el año 1280; el *DITECA*, formado por textos concejiles de Andalucía fechados entre el siglo XIII y el XVIII, o el *DiCCA XV*.

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

b) En lo diatópico y en distintos registros: el *CORDIAM*, con textos producidos únicamente en el continente americano entre 1492 y 1902 – hasta la fecha cuenta con 13868126 palabras en 18594 textos, distribuidos en diferentes géneros discursivos –; el *CODEA+ 2022*, con textos «desde la Cancillería hasta notas de manos inhábiles» (página inicial), de los que se ofrece la transcripción paleográfica, una presentación crítica y el facsímil, o el *ODE*, continuación del *CORDEREGRA*[2].

c) En el análisis de textos paralelos, como las distintas versiones de la Biblia, en *Biblia medieval*, corpus con unos cinco millones de palabras, proyecto dirigido por Andrés Enrique-Arias.

Como usuarios de estos y de otros corpus electrónicos, conscientes de lo complejo que resulta elaborarlos y prepararlos para su consulta, quisiéramos centrarnos en unos puntos que, con la ayuda de especialistas – ingenieros informáticos, filólogos, lingüistas e incluso historiadores –, sirvan en un futuro para ampliar y perfeccionar herramientas que mejoren su uso.

[5] Los corpus electrónicos sobre todo facilitan el trabajo a los lexicógrafos y a los expertos en lexicología, a quienes permiten señalar la procedencia de una voz y atestiguar su primera documentación, sus contextos de uso y su pérdida. Sin embargo, los investigadores centrados en la morfología, la sintaxis, la semántica o la pragmática históricas han de emplear mucho tiempo en discriminar los elementos que les interesan. Esta es la razón fundamental para señalar la necesidad de incrementar el número de lematizaciones, marcaciones y anotaciones de todo tipo que sirvan para sacar el máximo provecho en trabajos futuros.

2 Etiquetas

[6] Somos conscientes de que existen algunos programas de etiquetado, pero todavía dan un elevado número de errores, como comprueba Sanjurjo González (2017: 226-237) a través del software *TreeTagger*, y no abarcan todos los campos de estudio. Es deseable que un corpus histórico ofrezca el mayor número de anotaciones posible para identificar las características de cada texto, incluidas las referentes a las relaciones entre documentos. Lo ideal sería contar con etiquetas de distintos tipos para conseguir lematizaciones completas, por lo que profundizamos en la propuesta de Sierra Martínez (2017: 11 y ss.). Tomándola como base, hemos dividido en varios apartados nuestro listado de etiquetas, que habría que refinar

2 Algunos de los corpus mencionados en el texto forman parte de la red *CHARTA*, que fue creada con los objetivos de normativizar la presentación gráfica de los textos paleográficos y de establecer una metodología común para los documentos que los integran. Otros bancos de datos que la componen son el *Corpus de textos antiguos de Galicia* (*COTAGAL*), el *Corpus diacrónico de documentación malagueña* (*CODEMA*), el *Corpus de archivos privados de Navarra* (*CORAPRINA*), el *Corpus de documentos históricos de Mérida* (*CDHM*), el *Corpus de documentos de cancillería real* (*CODCAR*), la *Documentación de lamento en español desde orígenes* (*DOLEO*), etc. Remitimos a *CHARTA* (página *Subcorpus*).

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

antes de aplicarlas a los corpus históricos: filológicas (§ 2.1), discursivo-textuales (§ 2.2), gráfico-fonéticas (§ 2.3), morfosintácticas (§ 2.4), léxico-semánticas (§ 2.5) y sociopragmáticas (§ 2.6).

2.1 Etiquetas filológicas

[7] Para poder ubicar con precisión los cambios lingüísticos en el eje temporal, es necesario que los documentos incluidos en la base de datos sean lo más fiables posible tanto desde el punto de vista cronológico como desde el punto de vista filológico, por lo cual resulta imprescindible que los corpus marquen (i) si el texto es original o copia y (ii) si esta es coetánea o posterior, dado que las modernizaciones realizadas después por los copistas para adaptar el texto a la lengua de la época o a su dialecto particular pueden distorsionar la imagen de la etapa o de las etapas del cambio lingüístico que se esté investigando (Fernández-Ordóñez 2006: 1889). En este sentido, el *CORHEN* es uno de los pocos corpus que permite comprobar si el texto es original o copia y la fecha de esta.

[8] Lamentablemente, como pusieron de manifiesto Rodríguez Molina & Octavio de Toledo y Huerta (2017: 12-20), el *CORDE* no ofrece suficiente fiabilidad en lo relativo a la datación de ciertas obras medievales, como el *Cantar de mio Cid*, ya que mantiene la fecha de hacia 1140, propuesta por Menéndez Pidal (1976 [1944]), frente a la de 1200, propugnada, entre otros, por Montaner (1993), cuya edición, paradójicamente, utiliza dicho corpus, y no tiene en cuenta que la única copia disponible es del siglo XIV.

[9] Los trabajos que sostienen la fecha hipotética del siglo XII muestran una clara tergiversación de los datos, lo que produce resultados erróneos y, por consiguiente, lleva a manejar premisas falsas, como hemos comprobado en el proceso de gramaticalización de las construcciones verbales con *querer* + infinitivo (*quiere decir* 'significa'; *quiere llover* 'está a punto de llover', etc.). Si mantenemos la fecha defendida por Menéndez Pidal (1976 [1944]), podríamos pensar que la evolución de este verbo muestra el paso habitual de verbo léxico (*querer algo* / *a alguien*) a verbo auxiliar. Sin embargo, los datos demuestran que la dirección del proceso fue la opuesta: primero aparecen las construcciones verbales con *querer*, ya utilizadas en latín tardío; después, por desgramaticalización del auxiliar *querer* desde esos empleos construccionales – proceso motivado por la elipsis del infinitivo cuando este ha sido mencionado anteriormente en el discurso (cf. *Juan quiere comer un helado y María también quiere*) – se obtuvo el valor léxico (Zieliński, en prensa; Zieliński & Espinosa Elorza, en prensa).

[10] Como ilustramos en la tabla 1 siguiente, parece que hay un extraño aumento de frecuencia de usos léxicos en el siglo XII, pero, al eliminar los ejemplos del *Cantar de mio Cid*, se comprueba que los ejemplos con este valor, con un promedio del 17 %, están por debajo del construccional.

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

Tipo Siglo	Léxico	Elipsis	Construccional
XII	24 %	3 %	73 %
XIII	19 %	3 %	78 %
XIV	13 %	5 %	82 %
XV	22 %	3 %	75 %

Tabla 1. Distribución de *querer* según usos contextuales (Zieliński & Espinosa Elorza, en prensa)

[11] Por otra parte, se observan varios errores de datación y de autoría en algunas obras incluidas en el *CE* y en algunos textos procedentes de la base de datos *Electronic texts and concordances of the Madison corpus of early Spanish manuscripts and printings*, previamente elaborada por John O'Neill. En el *Poema de Fernán González*, por ejemplo, no aparece la fecha de redacción (figura 1).

Figura 1: Captura de pantalla del *CE*

Menosprecio de corte y alabanza de aldea se atribuye a John Stuart Mill y se da como fecha el año 1840 (1), cuando es bien sabido que el tratado fue escrito en 1539 por el moralista Antonio de Guevara, como se ve claramente en la parte final de la obra (2).

(1) Mill, *Menosprecio de corte y alabanza de aldea*, 1840, *CE*
Aquí se acava el libro llamado *Menosprecio de corte y alabanza de aldea*, compuesto por el ilustre señor don Antonio de Guevara …

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

(2) Antonio de Guevara, *Menosprecio de corte y alabanza de aldea*, 1539, *CORDE*
Aquí se acaba el libro llamado *Menosprescio de corte y alabanza de aldea*, compuesto por *el ilustre señor don Antonio de Guevara* […]. Fue impreso en la muy leal y muy noble villa de Valladolid por industria del honrado varón impresor de libros, Juan de Villaquirán a diez y ocho de junio, año de mil y quinientos y treinta y nueve.

La fiabilidad del *CE*, por lo menos en cuanto al texto seleccionado, queda anulada porque en la parte final del texto se ha acoplado el prólogo, que, a ojos de un investigador incauto, tergiversa por completo el contexto discursivo (ver figura 2).

Figura 2: Captura de pantalla del *CE*

[12] Respecto a la confección de los documentos, el *CORHEN* indica claramente el nombre de quien los escribió (3), anotación relevante para confirmar, entre otros detalles, las preferencias ortográficas de cada uno de ellos; por ejemplo, en el caso de los pronombres personales tónicos de primera persona, utilizan la grafía *hyo* Roi Sanchez, escribano público en Medina (4), Stephanus presbiter (5), Iohannis Sancii (6) y J. Filip, notario de la condesa donna Sancha (7). En el *CODEA+ 2022* también se han marcado los que denominan *copistas*.

(3) Doc. 0269, Briviesca, Burgos, 1196, *CORHEN*
Joh<a>n<ne>s de riolazedo scripsit

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

(4) Doc. 0203, San Salvador de Oña I, 1272, *CORHEN*
E por mayor seguramiento *hyo* hortun perez el sobredicho clérigo de Varanda

(5) Doc. 0369, Las Huelgas, *CORHEN*
hyo Roy ferrandez del Enbit de ea bona uoluntad vendo .ij. solares …

(6) Doc. 0426, Las Huelgas, 1237, *CORHEN*
Connusçuda cosa sea cuemo *hyo* don Johan peret …

(7) Doc. 0428, Las Huelgas, 1241, *CORHEN*
Que *hyo* dona. Vrraca por la gracia de dios Abbadessa de sancta Maria de Cannas …

[13] Es sumamente importante que los directores de corpus históricos seleccionen bien las ediciones críticas e incluyan también las paleográficas. Asimismo, debe ofrecerse, cuando se pueda, la consulta de facsímiles (García Moreno & Pueyo Mena 2017: 70). El *CORHEN* permite a los usuarios acceder tanto a la edición paleográfica como a la edición crítica de los documentos notariales seleccionados, mientras que el *CORDIAM* y el *CODEA+ 2022* ofrecen el facsímil de algunos documentos.

[14] En caso de que la obra sea una traducción de un texto foráneo, sería muy útil incluir el texto original o, por lo menos, la información de cómo acceder a él. Asimismo, se ha comprobado en varias publicaciones la utilidad del empleo de textos paralelos, como las distintas versiones de la Biblia (Enrique-Arias 2012: 85).

[15] Como el cambio lingüístico no solo se produce en el eje temporal, sino también en el eje espacial, a través del cual se difunde y se propaga, resulta de gran ayuda la inclusión de la marcación tópica. Si bien tanto el *CORDE* como el *CORDIAM* permiten a sus usuarios delimitar la búsqueda por países, sería recomendable ahondar más en este aspecto para deslindar zonas en los textos medievales documentales. El *CORHEN* indica el centro donde se redactaron o recopilaron y la localidad que se menciona en el texto o la supuesta, obtenida gracias a la información que manejan los transcriptores. Es bien sabido que los copistas y amanuenses no solo transformaban el texto según el modelo de la lengua que tenían ante sus ojos, sino que lo adaptaban, a menudo, a su propia modalidad lingüística, «sin sentir que estaban atravesando una frontera» (Fernández-Ordoñez 2006: 1792). Esto permite explicar con más precisión la difusión de una innovación lingüística desde un determinado centro monacal (Wright 1989 [1982]: 165).

[16] Los datos de corpus permiten también comprobar la extensión de los tratamientos. El de *señor* (< lat. *seniore*) en suelo hispano se corresponde con la lenta imposición del sistema feudal en los reinos norteños (figura 3).

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

Figura 3: Extensión del tratamiento *señor* en el primitivo iberorromance (Zieliński 2021: 39)

Se introduce por la Marca Hispánica, desde donde empieza a difundirse a la zona navarra, a la aragonesa y, en parte, a la castellana, de tal manera que en el siglo XI se observa una isoglosa, ubicada a la altura del río Pisuerga, que divide los primitivos romances hispánicos en occidentales, con el tratamiento vernáculo procedente de lat. *dominus* (8), (9) y (10), y orientales (11), (12) y (13), con el nuevo tratamiento procedente de lat. *senior*, reservado en esta fase a referentes seculares (Zieliński 2021: 37-40).

(8) Anónimo, Alfonso VII da a Gutierre y a doña Toda su mujer, la villa de Foramnada, León, 1149, *CORDE*
Ego igitur Adefonsus, tocius Ispanie imperator, una cum filiis meis rege *domno* Sancio atque Fredinando, facio cartam donacionis vobis domno Guteirrio et uxori vestre *domine* Tote de una villa …

(9) Anónimo, Alfonso VII da a Gutierre y a doña Toda su mujer, la villa de Foramnada, León, 1149, *CORDE*
Contrario dicebant comes Rodericus et frater ejus Fredenandus Didaz quod fuerat supra memoratum monasterium de Taule ex eorum progenie et ipsi debebant illud habere post partem matris sue *domne* Christine et amite sue *domne* Urrace comitisse …

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

(10) Anónimo, Fueros de la villa de Palenzuela, 1074, *CORDE*
Si ille *dominus* qui mandaverit Palenciolam Comitis voluerit enviare in mandaderia militem aut pedonem de Palenciola, det ei totam suam espensam …

(11) Anónimo, Sancho Garcés II Abarca y la reina Urraca confirman al monasterio de Pampaneto la villa de Senzano, ca. 985, *CORDE*
Eximino *senior* confirmat, Abeiz *senior* confirmat …

(12) Anónimo, Escritura dotal en el matrimonio de Ramón Berenguer I el viejo con Elisabet, ca. 1039 (Mateu Ibars & Mateu Ibars 1980-1991: 448)
Ego Raimundos gracia dei comes una per voluntantem dei atque *Seniorum* electione expetiu in matirmonio nomine Elisabet gracia dei comitissa…

(13) Anónimo, Carta de población de Cardona, 986, *CORDE*
seu pontifices, seu clericorum, abbatum, monachorum, et omnem gradum Ecclesie, sive laicos, vices commites et *seniores*, vel viliores personas regimini nostro parencium …

2.2 Etiquetas discursivo-textuales

[17] Una innovación lingüística también se consolida a través de una determinada tradición, de ahí que sea esencial que en los corpus históricos se integre la marcación de este condicionante del cambio lingüístico que fija, en cierta medida, el empleo de determinados elementos o de ciertas expresiones lingüísticas (Kabatek 2005: 159-160).

[18] Es cierto que el *CORDE* permite, en principio, delimitar la búsqueda a través de la marcación de unos 170 criterios diferentes, calificados por la Real Academia Española de *temas*, término demasiado vago desde el punto de vista discursivo, ya que varios pueden concurrir en una misma tradición discursiva. Es el caso del legado historiográfico atribuido al rey Alfonso X, considerado como *historia y documentos*; los documentos, aunque aluden a acontecimientos históricos, remiten incuestionablemente a una tradición discursiva diferente. Nótese, además, que el etiquetador inserto en el buscador del corpus académico no faculta el rastreo por las distintas etiquetas establecidas. A este respecto, en el *CORDIAM* se accede con más facilidad a la localización de los elementos en búsqueda a través de tres tipos textuales: (i) documentos – administrativos, cronísticos, jurídicos y particulares (cartas) –, (ii) literatura – narrativa, poesía, prosa varia, teatro y textos cronísticos – y (iii) prensa – informativos, comentarios y publicitarios –. Otros corpus, como el *CORHEN*, se ciñen a textos documentales medievales.

[19] Incluso se podría anotar la parte del texto en la que aparece el término buscado: una glosa, una cita, el encabezado, la dedicatoria, etc., que, si bien cons-

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

tituyen una tradición discursiva aparte, suelen aparecer conjuntamente en el macrocontexto del elemento que nos interesa, como se observa en (14). También resulta interesante (15), ejemplo que en los metadatos señala a qué tipo de texto – dedicatoria – corresponde el fragmento seleccionado.

(14) Barahona de Soto, *Las lágrimas de Angélica*, 1586, *CORDE*
Fecha en Tous, a XXI días del mes de junio de mil y quinientos e ochenta e cinco años. Yo el Rey. Por mandato de su Majestad, Antonio de Eraso. *Al excelentísimo señor Don Pedro Girón, Duque de Osuna, Conde de Ureña y Virrey de Nápoles*.
Excelentísimo señor: Estos doce cantos …

(15) Balbuena, *Grandeza mexicana*, 1604, *CORDIAM*
Al ilustrísimo y reverendísimo Don Fray García de Mendoza y Zúñiga. Arzobispo de México, del Consejo de su Majestad, etcétera\\\ // Habiendo amagado a escrebir estas excelencias de México con deseo de darlas a conocer al mundo viéndolas hoy aumentadas y en todo su colmo y lleno con la deseada venida de Vuestra Señoría Reverendísima, paréceme que no cumpliera con lo que a ellas y a mis deseos debo si a todos juntos no hiciera un nuevo servicio …

Sería deseable especificar, como indica Sanjurjo González (2017: 237), el «conjunto de estructuras internas de un texto que hacen que sea reconocible como miembro de un género textual concreto», pero hasta el momento «no hay unanimidad a la hora de establecer etiquetas […] estándar».

2.3 Etiquetas gráfico-fonéticas

[20] Con base en el corpus *Xiga*, compilado por el Seminario de lingüística informática de la Universidad de Vigo, con textos sobre informática y telecomunicaciones escritos en gallego – algunos con normas ortográficas distintas a la oficial –, Aguirre Moreno, Andión Rodríguez & Gómez Guinovart (2001: 5) utilizan un etiquetado con aclaraciones pertinentes, como la equivalencia gráfica a la forma normativa, los *lapsus calami* y los usos por desconocimiento de la norma, marcas que les parecen insuficientes.

[21] Sería preciso etiquetar, sin agotar las posibilidades:

a) Las pausas. En algunas ediciones paleográficas, como las que ofrece el *CORHEN*, sin lematizar, se puede ver la marcación de pausas, fundamental para la confirmación de que ciertos elementos o expresiones funcionan, por ejemplo, como marcadores del discurso (16). Para ciertos tipos de topicalización no son tan necesarias, ya que la pista está en el empleo del pronombre reasuntivo (17).

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

(16) Doc. 0241, San Salvador de Oña I, 1279, *CORHEN*
E demas desto; otorgamos & uenimos de connosçido que Reçebimos de uos dos mil morauedis …

(17) Doc. 0478, Las Huelgas, Burgos, 1288, *CORHEN*
todo uos *lo* do por uuestro …

b) La separación de los párrafos. Muchos textos medievales presentan signos distintivos, tomados, como bien saben los paleógrafos, de los empleados en las partituras.

c) La división de palabras, como en los adverbios en *-mente* (18), y, en el caso opuesto (19), las contracciones, caso de, por ejemplo, las amalgamas de preposición y artículo (20).

(18) Doc. 0246, San Salvador de Oña I, 1279, *CORHEN*
que nos y auemos todo *entera mente* …

(19) Doc. 0143, San Salvador de Oña I, 1238, *CORHEN*
E fazet la lauor bien e *lealmientre* …

(20) Doc. 0025, San Salvador de Oña I, 1102, *CORHEN*
Ennos barrios quantum ibihabemus [sic] …

d) La presencia o ausencia de tilde en la misma palabra para agilizar las búsquedas. En la forma verbal del futuro del verbo *ser*, el *CORHEN* solo emplea la forma sin tilde (21); sin embargo, en el *CORDE* aparecen las dos opciones (22, 23).

(21) Doc. 0143, San Salvador de Oña I, 1238, *CORHEN*
e alos ke *seran* siempre hi moradores …

(22) Anónimo, *Fuero de Viguera y Val de Funes*, ca. 1130, *CORDE*
E los VI deslindadores, *seran* de su gent …

(23) Anónimo, *Pleito entre el abad de Valbona y el concejo de Velosiello por unos pastos y unos montes*, 1208, *CORDE*
Conocida cosa sea a los que son e a los que *serán* que …

e) El uso de un mismo término con mayúscula o con minúscula. En el *CORHEN*, aunque se busque con mayúscula el sustantivo *Villa* en distintos topónimos, se ofrecen también los casos con minúscula (24). En la edición crítica se normaliza la grafía (25).

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

(24) Doc. 0139, San Salvador de Oña I, 1231, *CORHEN*
apud *villam* y mara

(25) Doc. 0139, San Salvador de Oña I, 1231, *CORHEN*
apud *Villam* Imara

f) El uso de un mismo término con *u*, *v* o *b* (26, 27).

(26) Doc. 0142, San Salvador de Oña I, 1236, *CORHEN*
la quarta part. de todo quanto yo & mi mugier donna Vrraca *auemos* …

(27) Doc. 0277, Las Huelgas I, 1200, *CORHEN*
vendemos la nuestra part del azenia que *abemos* conbusco …

g) El uso de un mismo término con *z*, con *ç* y con *c*. Lo ejemplificamos con distintas formas del verbo *fazer* (28-30).

(28) Doc. 0158, San Salvador de Oña I, 1250, *CORHEN*
ante ke metades la foz en las miesses *fazet* nos lo saber …

(29) Doc. 0477, Las Huelgas, Burgos, 1288, *CORHEN*
que el Rey don Sancho nuestro ssennor mando *ffaçer* …

(30) Anónimo, *Cortes de Benavente*, 1202, *CORDE*
faga della tal fuero qual *facen* las otras heredades de los Cavalleros

h) El uso de un mismo término con *f-*, con *h-* o sin grafía inicial (31-33).

(31) Doc. 0419, Las Huelgas, 1233, *CORHEN*
& Casas & *forno* del Abbatissa de villa Mayor …

(32) Anónimo, *Carta de donación, documentos de Alfonso X dirigidos a Andalucía*, 1253, *CORDE*
Et este *horno* sobredicho vos do e vos otorgo …

(33) Anónimo, *Carta de censo, colección diplomática de Santo Toribio de Liébana*, 1463, *CORDE*
que tengades el dicho solar poblado por sienpre jamas e lo levedes con el dicho *orno* e molino …

Quedan los problemas con las sibilantes, términos con diversas manifestaciones de los resultados de las vocales latinas *ĕ* y *ŏ* tónicas y un largo etcétera.

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

2.4 Etiquetas morfosintácticas

[22] Disponemos de etiquetadores morfosintácticos, como el *Part-of-speech tagging* o *POS-tagging*, mediante el cual a cada palabra de un texto se le asigna una categoría funcional y alguna información adicional, como subcategorías o lemas asociados (Oliver González 2021). Sirva de ejemplo el lexema *casa*, que puede ser o bien sustantivo o bien verbo.

[23] Oliver González (2021) menciona una propuesta de etiquetario morfosintáctico universal: el *universal target*, que incluye ADJ, ADP, ADV, CONJ, DET, NOUN, NUM, PRT, PRON, VERB. También distingue: (i) el etiquetador por unigramas, que solo tiene en cuenta una palabra; (ii) el etiquetador por bigramas, que contempla también la palabra anterior, y (iii) el etiquetador por trigramas, que se amplía a las dos palabras anteriores. En su opinión, se deberían utilizar corpus etiquetados manualmente.

[24] El etiquetador *Thera*, del Centre de Llenguatge i Computacio de la Universitat de Barcelona, cuenta con más de veinte años de experiencia en el análisis morfológico como lematizador, flexionador y etiquetador. También *FreeLing*, biblioteca que ofrece servicios como el análisis morfológico, el tratamiento de sufijos o el reconocimiento de varias palabras, permite detectar pronombres enclíticos, formas verbales y sufijos diminutivos y aumentativos de sustantivos y adjetivos (Vilar Bohigues 2010: 5).

[25] Ribera, Molina & Pla (2000) hablan de un analizador sintáctico parcial basado en expresiones regulares, del analizador morfológico *MACO* y del analizador parcial de oraciones en lenguaje natural *APOLN*. Asimismo, especifican las etiquetas léxicas *PAROLE*: núcleos nominales y verbales, sintagmas nominales, sintagmas preposicionales y sintagmas verbales.

[26] Para Graña Gil (2000: 215-219), resulta indispensable el contexto y aconseja sistemas como *TnT – Trigrams'n'Tags* – (Brants 1996), entre otros, que ofrecen los mejores rendimientos siempre y cuando los textos lleguen al sistema de etiquetación correctamente segmentados; si no, son frecuentes las ambigüedades al tratar elementos como *ténselo*, formado por el verbo *tener* más dos pronombres o por el verbo *tensar* más un pronombre; o como *sin embargo*, expresión adverbial o preposición más sustantivo, lo que prueba que «las técnicas no han alcanzado todavía el grado de madurez deseable»[3] (Graña Gil 2000: 219).

3 Aguirre Moreno, Andión Rodríguez & Gómez Guinovart (2001) adoptan el sistema de marcación *Text encoding initiative* (*TEI*) y la propuesta del *Expert advisory group on language engineering standards* (*EAGLES*) en lo relativo a las categorías gramaticales y rasgos morfosintácticos en el corpus *Xiga*, con textos escritos en gallego. Distinguen las categorías de nombres comunes y propios, con género y número; adjetivos; artículos; pronombres: formas rectas, oblicuas, ligadas y átonas; contracciones; verbo: modo, tiempo, número y persona, formas personales e impersonales; interjecciones, conjunciones y preposiciones; locuciones prepositivas, conjuntivas y adverbiales; puntuación, y *residual*, novedad que toman de *EAGLES*: palabra extranjera, fórmula, símbolo, acrónimo, abreviatura y sin clasificar. A partir de los criterios de la red *CHARTA*, *TEI*, por ejemplo, en el *COHREN*, permite a los usuarios recuperar estructuras sintácticas complejas (Isasi Martínez et

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

[27] En el proyecto del *CODEA+ 2022*, desde el que se ha originado el *Atlas lingüístico diacrónico y dinámico del español*, se pueden buscar, «aparte de las distribuciones léxicas, la extensión geográfica de elementos morfosintácticos como [*otro*/*otri*/*otre*/*otrie*], *agora*/*ahora*, **mentre*/**mientre*, el superlativo en [**ísimo*], formas verbales en [**rá*/**drá*], colocaciones [*no … ning**], [*no … alg**] y cualquiera que se le ocurra al usuario» (*CODEA+ 2022*, página inicial).

[28] El *CORHEN*, aunque no está lematizado, en *LYNEAL*, diseñado por Hiroto Ueda, permite buscar la estructura artículo + posesivo, todas las formas verbales de verbos irregulares: *compr{ar}* lleva a *compro*, *compré*, *compramos* …, o colocaciones, como *ningun*/*a*/*os*/*as* + una o dos palabras + *non*.

[29] Desde nuestro punto de vista, para un corpus histórico sería de gran utilidad etiquetar:

a) La clase de palabra, para discriminar términos coincidentes en la forma, como *cabe*, preposición (34) y verbo (35).

(34) Doc. 0434, Las Huelgas, 1243, *CORHEN*
Reçebemos de uos .j. tierra en Ormaza *cabe*l nuestro Moljno de la vega …

(35) Anónimo, *Fuero de Brihuega*, ca. 1242, *CORDE*
tornes a cada uno de los fiadores por quantol *cabe* en la debda …

b) El género, como en el sustantivo *pro* (36, 37), femenino y masculino, y el número en *uebos* (38), singular acabado en *-s* (< lat. *opus*). En el *CORHEN* se permiten búsquedas en singular y plural.

(36) Doc. 0184, San Sanvador de Oña I, 1265, *CORHEN*
& que ffagades hy *la pro* que podierdes …

(37) Anónimo, carta de concesión, Documentos del Reino de Castilla, 1228, *CORDE*
& que fagades en ela *el pro* que pudierdes …

(38) Anónimo, documento de avenencia, Documentos del Reino de Castilla, 1228-1232, *CORDE*
& dixieron lo que es sobrescripto, que auien antes dicho, del mondar de los calzes & del auentadero quando *uebos* fuesse …

c) Morfemas flexivos (39-41), para marcar, por ejemplo, la historia de la desinencia de segunda persona del plural, y derivativos, tanto prefijos (42, 43) como sufijos (44).

al. 2020).

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

(39) Doc. 0062, San Salvador de Oña I, 1188, *CORHEN*
Et uos Guter pelaez *dades* ad nos pro ipsa terra …

(40) Juan de Tapia, sin título, *Cancionero de Estúñiga*, ca. 1407-1463, *CORDE*
Pues la muerte *daes* a uos, / la uida se nos destierra / …

(41) Anónimo, *Libro del caballero Cifar*, 1300-1305, *CORDE*
– Señor, ¿a quién nos *dais* por capitán?

(42) Doc. 0444, Las Huelgas, 1251, *CORHEN*
E por estas vinnas *sobrescriptas* que nos damos auos …

(43) Doc. 0179, San Salvador de Oña I, *CORHEN*
Que por esta sententianon *menoscaben* cosa ninguna de su derecho …

(44) Doc. 0215, San Salvador de Oña I, 1275, *CORHEN*
que dedes cadanno al abbat vna *procuraçion* bien & cunplida mientre a el & a toda la conpanna que fuere con el …

d) Compuestos, tanto en elementos léxicos (45) como gramaticales (46, 47).

(45) Doc. 0151, San Salvador de Oña I, 1245, *CORHEN*
Desto son testes. Peidro armilloz *fidalgo* testis. Gonzaluo royz *fidalgo* testis …

(46) Anónimo, *Tratado de plantar o enjerir árboles o de conservar el vino*, 1385-1407, *CORDE*
E *assi mesmo* los nabos seran mucho dulces …

(47) Almerich, *La fazienda de Ultra Mar*, ca. 1200, *CORDE*
sobre el .iiii.º non tornaré *porque* vendieron el iusto por argent e el mezquino por calçado …

e) Fusiones de constituyentes, como pronombres enclíticos con el verbo (48), y contracciones (49).

(48) Doc. 0233, San Salvador de Oña I, 1278, *CORHEN*
Et *diogelo* que lo ayan libe & quito …

(49) Doc. 0138, San Salvador de Oña I, ca. 1229, *CORHEN*
sediendo *enna* casa …

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

f) Colocaciones, entendidas como «combinaciones léxicamente restringidas de dos unidades léxicas: una que el hablante escoge libremente para expresar sus necesidades comunicativas – la base de la colocación – y otra – el colocativo – seleccionada de manera léxicamente restringida en función a la base para expresar un sentido particular» (Alba-Salas 2012: 5), como *buscar achaque* 'poner excusa' (50) o *poner culpa* 'responsabilizar de una acción' (51). Por el momento, únicamente el *CODEA+ 2022* ofrece una pestaña exclusiva para buscar colocaciones.

(50) Alfonso X, *Estoria de España II*, 1270-1284, *CORDE*
et non quisieron estonce descubrir sus coraçones & *buscaron achaque* para salir se del palacio …

(51) Anónimo, *Libro de los buenos proverbios que dijeron los filósofos y sabios antiguos*, ca. 1250, *CORDE*
Non *pongas culpa* a Dios en yerro que tu ffagas …

g) Orden de palabras. Por mencionar solo una posibilidad, el etiquetado sería muy útil para discriminar tópicos o focos, o para observar el cambio de colocación de los clíticos antes y después del siglo XIV. Hasta esa centuria, en estructuras con tópico – que subrayamos –, se prefiere la posposición del clítico al verbo (52); después es más frecuente la anteposición (53) (Fernández-Ordóñez 2009: 147).

(52) Alfonso X, *General Estoria. Primera parte*, ca. 1275, *CORDE*
E aun esso que dava faziél*o* de mala voluntad …

(53) Martínez de Toledo, *Corbacho*, 1438, *CORDE*
pero al coraçón espiritual non *lo* puede temtar …

h) Ubicación y valores en contexto del adverbio de negación y de las conjunciones copulativas o disyuntivas. Los corpus únicamente posibilitan el cómputo de las unidades halladas y, gracias al operador de distancia, podemos rastrear la aparición de dos términos o solo de uno, a una distancia concreta de uno de ellos, pero el investigador se ve obligado a discriminar miles de casos para analizarlos. Obsérvese la diferencia de la negación oracional y de la negación de constituyentes en (54) y (55), respectivamente, o la negación expletiva (56) (Camus Bergareche 2006: § 13.2.1.1 y 13.2.3.2), y usos no copulativos de la conjunción *e*(*t*)/*y* cuando no suma, esto es, con valor ilativo (57), e incluso cuando funciona como marcador de inicio de turno de habla (58) (Garachana 2014: § 21.9.2 y 21.9.3).

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

(54) Doc. 0500, Aguilar de Campoo, Palencia, 1113, *CORHEN*
Et si aliquis homo istam scripturam ad disrumpendum venerit vel ad iudicium compulsaverit, *non* habeat parte cum Deus nisi cum Iudas traditore …

(55) Doc. 0530, Barrio Panizares, Burgos, 1196, *CORHEN*
non p<er> metu<m> neq<ue> p<er>turbatu<m> sensu<m> sed p<er> spontaneas n<ost>ras uoluntates uendemos ad uos frepetro amigo de t<er>radiellos illo molino que habem<us> e<n>na molina depanizares …

(56) Doc. 0610, Aguilar de Campoo, Palencia, 1223, *CORHEN*
E fre Petro el Negro tenía a Cordovilla e vedónos la villa, que *non* entrássemos en ella…

(57) Doc. 0654, La Vid de Ojeda, Palencia, 1243, *CORHEN*
Et porque este mandamiento que nós mandamos sea más firme, nós devandichos don Martín prior de Fusiellos, e don García Roíz Sarmiento e Pedro Roíz Calderón mandamos fazer tres cartas partidas por ABC …

(58) Anónimo, *Calila e Dimna*, 1251, *CORDE*
Dixo Sençeba: – ¿*Et* qué es eso?

La lista de etiquetas sintácticas sería innumerable y nuestro espacio es reducido, de ahí que remitamos al lector a las consideraciones integradas en los distintos capítulos de las tres partes hasta ahora publicadas de la *Sintaxis histórica de la lengua española* (Company Company 2006, 2009, 2014).

2.5 Etiquetas léxico-semánticas

[30] Aunque hayamos afirmado que los corpus ayudan especialmente a los investigadores en léxico, su labor no está carente de obstáculos, por lo que ayudaría una elaboración cada vez más amplia y precisa de etiquetas de todo tipo en este ámbito. Ya disponemos de algunos etiquetados léxico-semánticos: Aguirre Moreno, Andión Rodríguez & Gómez Guinovart (2001: 6) aplican al corpus *Xiga* las etiquetas *NAME*, que incluyen el atributo *type*, que especifica su clase semántica. También para el gallego existe *SLI-Tagger*, etiquetador morfosintáctico y semántico del gallego. Por su parte, Montoyo (2002) abordó la desambiguación léxica mediante marcas de especificidad, y Suárez Cueto (2004) y Nica (2004) trataron la desambiguación semántica automática. Posteriormente, Goded Rambaud & Ibáñez Moreno (2012) aplicaron un etiquetado semántico-cognitivo a un corpus más concreto: el de notas de cata de vino en español.

[31] Para el catalán y el español, Taulé Delor et al. (2006) mencionan corpus etiquetados sintáctica y semánticamente, sobre todo para marcar el número de argumentos exigidos por el predicado verbal y el tipo de papel temático. Se ciñen a

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

los verbos de estado, de actividad y de realización, que dan lugar a una guía de etiquetas automáticas de papeles semánticos.

[32] Fernández Reyes, Leyva Pérez & Lau Fernández (2011: 55-64) ofrecen algunas consideraciones para el diseño de una herramienta de análisis semántico que solvente la ambigüedad en el sentido de las palabras, «problema actual, aún sin resolverse completamente». Toman el ejemplo de *Yo toco el bajo en los bajos de la escalera*, en el que *bajo* puede referirse a una parte del pantalón o a un instrumento musical, y *bajos* puede ser, en su opinión, un adverbio de lugar – apreciación errónea – o un sustantivo. Utilizan *WordNet*, base de datos léxica que tiene en cuenta las categorías de sustantivo, verbo, adjetivo y adverbio, y las relaciones semánticas de hiponimia, hiperonimia, meronimia, holonimia, sinonimia, antonimia y troponimia, detalles más útiles para la elaboración de diccionarios en red que para corpus digitalizados.

[33] Con vistas a un corpus electrónico histórico, sería de gran provecho la marcación del significado de un vocablo en un determinado contexto[4]:

a) Homonimia parcial. Afecta a elementos que presentan homofonía y homografía, pero pertenecen a distintas categorías, como la preposición *so* (59), procedente de lat. *sub*, y el posesivo *so*, originado en lat. *suu*(*m*), con sustantivos masculinos (60) y femeninos (61).

(59) Doc. 0161, San Salvador de Oña I, 1254, *CORHEN*
ni lo ayades poder de uender ni de empennar, ni de meter *so* otro sennorio …

(60) Doc. 0161, San Salvador de Oña I, 1254, *CORHEN*
& aell anno coja *so* fructu …

(61) Doc. 0142, San Salvador de Oña I, 1236, *CORHEN*
adonna Vrraca Roiz. *so* mugier de don Orti Ortiz …

b) Homonima absoluta. Los términos presentan homofonía y homografía, y pertenecen a la misma categoría, como *llama* 'lengua de fuego' (62), del lat. *flamma*, y *llama* 'variedad doméstica del guanaco' (63), del quechua.

4 Dado que la monosemia es prácticamente imposible de encontrar, se podrían marcar los términos con un solo significado. Podríamos pensar en *casa* (i), pero, en realidad, tiene más, como 'ciudad' cuando se emplea como arabismo semántico (ii):

(i) Doc. 0166, San Salvador de Oña I, 1257, *CORHEN*
& estableçemos que nos fagades hy una *casa* de .vj. braças…

(ii) Anónimo, *Poema de mio Cid*, ca. 1140, *CORDE*
Vedada l'an conpra dentro en Burgos la *casa*

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

(62) Alfonso X, *Lapidario*, ca. 1250, *CORDE*
Et quando la queman faze *llama* & sal della fumo …

(63) Gonzalo Fernández de Oviedo, *Historia general y natural de las Indias*, 1535-1557, *CORDE*
En la tierra llana llaman a este animal col, e en la sierra le dicen *llama* …

c) Polisemia. El sustantivo *orden* puede presentar, entre muchos otros, los significados de 'grupos de monjes que siguen una regla' (64) o 'hábito' (65).

(64) Gonzalo Fernández de Oviedo, *Historia general y natural de las Indias*, 1535-1557, *CORDE*
quando entraron en la *orden* …

(65) Doc. 0432, Las Huelgas, 1243, *CORHEN*
e dieron le la *orden* …

Desde el punto de vista estrictamente léxico, sería interesante la discriminación de términos cultos (66), semicultos (67, 68) y populares (69) tanto en la lengua escrita como en los registros orales; los cruces léxicos o fusiones, como *patata*, de *papa* y *batata*; o *diomingo*, de *día* y *domingo* (70-74), y las expresiones fraseológicas, como *a furto* 'a hurtadillas' o *de cabo a cabo* 'completamente' (75).

(66) Doc. 0418, Las Huelgas, 1233, *CORHEN*
que cante *Missa* por mj *anima* cada dia …

(67) Doc. 0171, San Salvador de Oña I, 1259, *CORHEN*
por la *graçia* de dios Abbat de Onna …

(68) Doc. 0171, San Salvador de Oña I, 1259, *CORHEN*
& sobre este *pleyto* rogaron nos por el Rey que …

(69) Doc. 0171, San Salvador de Oña I, 1259, *CORHEN*
que les den sos *pennos* …

(70) Hernán Cortés, *Cartas de relación*, 1519-1526, *CORDE*
y *patata* yuca, así como la que comen en la isla de Cuba …

(71) Francisco Pizarro, *Instrucción impartida por Francisco Pizarro y fray Vicente de Valverde a Diego Verdejo*, 1540, *CORDE*
y si es tierra de mahiz o de *papas* …

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

(72) Gonzalo Fernández de Oviedo, *Historia general y natural de las Indias*, 1535-1557, *CORDE*
salvo que la *batata* es más delicada fructa o manjar …

(73) Doc. 481,1277, *CARRIZO*
Fecha esta carta en Mirauales, *diomingo*, .xviii. dias de abril …

(74) Anónimo, *Carta del rey don Alfonso*, 1273, *CORDE*
o que la venden *a furto* …

(75) Anónimo, *Fuero General de Navarra*, 1250-1300, *CORDE*
no aylenando las vinas podando & cauando todas *de cabo a cabo* …

2.6 Etiquetas sociopragmáticas

[34] El auge de la pragmática histórica o de la sociopragmática histórica que se observa desde los últimos años representa un verdadero desafío a la operatividad de los actuales corpus históricos, que tienden a orientar la investigación tradicional con un «corpus-based approach» (McEnery, Xiao & Tono 2006: 8-10), ya que (i) es muy difícil encontrar textos antiguos que remitan estrictamente al registro oral – recuérdese que las primeras obras teatrales españolas, que representan por mimesis el lenguaje oral, datan del siglo XVI y también de esa centuria proceden los primeros textos conservados del género epistolar –; (ii) las unidades pragmáticas, independientemente de su naturaleza, se caracterizan por su elevada polisemia funcional (por ejemplo, *mande*, *anda*, *beso las manos de* …, etc.) y, por lo tanto, (iii) para analizarlas debemos disponer de un contexto mucho más amplio que el que ofrece un corpus histórico tradicional. Si fuera posible, se deberían incluir todas las variantes de naturaleza sociolingüística, históricamente fluctuantes, como quién lo enuncia, a quién lo dice, en qué situación comunicativa y psicológica lo enuncia, y cuál es la fuerza ilocutiva del acto de habla analizado (Brinton 2012: 107-110).

[35] El primero de los problemas se ha resuelto parcialmente gracias a la aportación de Koch & Oesterreicher (1985: 20-25). Sabemos que en ciertos géneros discursivos hay pasajes dialogados que, aunque planeados, codifican lo que Zumthor (2001 [1987]: 19) denomina *oralidad mixta*, como se ve en (76) y (77), ejemplos que reflejan elementos dialogados. Con la ayuda de un buen etiquetado se podría restringir su búsqueda.

(76) Anónimo, *Libro del cavallero Çifar*, 1300-1305, *CORDE*
el ribaldo se fue allí do los cavalleros peleavan &, quando fue cerca dellos, conosció el cavallero Cifar en las vestiduras que le avía dado & díxole: –

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

Amigo, ¿aquí estás? Tú seas *bien venido*. – Señor – dixo el ribaldo – aquí estó a vuestro servicio …

(77) Anónimo, *Cantar de Mío Cid*, h. 1140, *CORDE*
cuando llegó Avengalvón, dont a ojo lo ha, / sonrisándose de la boca ívalo a abraçar, / en el ombro lo saluda, ca tal es su usaje: / – ¡Tan *buen día* convusco, Minaya Álbar Fáñez!

El segundo obstáculo es más difícil de superar. Según el estado de las herramientas informáticas actuales, no puede resolverse a través de ningún etiquetado previamente elaborado, porque, como indican McEnery, Xiao & Tono (2006: 106), el significado pragmático no puede detectarse automáticamente debido a su polisemia funcional. Si buscamos la interjección *mande*, utilizada para dos fines pragmáticos – (i) para responder al llamamiento del destinatario y (ii) para pedirle que repita el enunciado que el emisor por alguna razón no ha entendido –, en una época concreta, por ejemplo 1750-1800, el *CORDE* ofrece 240 ejemplos con todos los valores, procedentes de 116 textos elaborados en España, que requieren gran cantidad de tiempo de análisis por parte del investigador para discriminar la función que interesa. Por lo tanto, la descodificación de la unidad pragmática buscada es subjetiva. Esto explica por qué las investigaciones sobre pragmática histórica que utilizan los corpus históricos aplican, ante todo, una aproximación basada en el uso (Brinton 2012: 107-110).

[36] La tercera y última dificultad es particularmente importante para los estudios dedicados a la (des)cortesía verbal desde el punto de vista diacrónico, dado que los términos o expresiones analizables requieren un contexto amplio, incluso un texto completo si se trata, por ejemplo, del género epistolar, para comprobar el carácter de las relaciones sociales existentes entre el remitente y el destinatario. Asimismo, debe incluir todas las variantes sociolingüísticas posibles, que, al inscribirse en el índice de contextualización de Gumperz (1982: 132), condicionan el empleo de una unidad pragmática en una situación comunicativa dada. En este sentido, el *CORDIAM* permite, en principio, delimitar la búsqueda a través del etiquetado de sexo (hombre, mujer, varias) y de raza, aunque esta posibilidad se circunscribe a los documentos. También, *CODEA+ 2022* permite discriminar la búsqueda por sexo, delimitando la voz femenina al papel de emisor, destinatario o escribiente/firmante. En cuanto al *CORHEN*, normalmente se puede encontrar el dato de quién escribe el documento o, al menos, quién lo manda redactar.

[37] Ahora bien, a la hora de diseñar un corpus histórico para fines pragmáticos, sería necesario tomar en cuenta el trabajo de historiadores, especializados en las relaciones socioculturales de las épocas pasadas, que puedan justificar el etiquetado relativo a la posición social de los participantes del acto de habla concreto que analicemos. Todo esto es especialmente deseable para estudiar en profundidad los mecanismos pragmáticos del español áureo, época en la que las relaciones so-

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

ciales de la estancada sociedad del Antiguo Régimen resultan sumamente complejas a ojos de un filólogo incauto (Domínguez Ortiz 2012 [1973]: 69), ya que con frecuencia no encajan en la diada de poder-solidaridad, mencionada con asiduidad a partir del estudio de Brown y Gilman (1960: 257 y ss.). Su aplicación a los corpus históricos contribuiría a entender mejor los mecanismos pragmáticos de cortesía volitiva que caracterizan el sociolecto del estamento más privilegiado (Moreno 2002: 32-35).

3 Conclusiones

[38] Esta tarea de etiquetado, tan compleja, minuciosa y dilatada en el tiempo, se podrá abordar – esperemos que en un futuro próximo – siempre que se formen equipos interdisciplinarios e interuniversitarios[5], incluidos los de centros de investigación superior, que encuentren financiación no solamente a través de proyectos de investigación estatales o autonómicos, sino también mediante contratos universidad y centros de investigación superior-empresa.

[39] Se podría comenzar por corpus específicos con un número no muy elevado de textos que pertenezcan a un solo género discursivo para que el léxico empleado no sea demasiado variado, como el *CORHEN*, o por una o varias partes de corpus más amplios, con el objetivo de confeccionar el etiquetado en todos los niveles que se han mencionado en este trabajo y facilitar su aplicación a otras partes y a otros corpus.

[40] En estos equipos se necesita personal especializado que supervise de principio a final el proceso de etiquetado, ya que este debe ser «constantemente un producto actualizable y reutilizable» (Llisterri Boix & Torruella Casañas 1999: 46).

5 Buen ejemplo del sincretismo interdisciplinar que proponemos lo constituye, por poner tan solo un ejemplo, el banco de datos *HESPERIA*.

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

Abreviaturas y referencias bibliográficas

ADMyTE = Francisco Marcos Marín, Charles B. Faulhaber, Ángel Gómez Moreno, Antonio Cortijo Ocaña (eds.) 1991-. *Archivo digital de manuscritos y textos españoles*. https://www.admyte.com.

Aguirre Moreno, Andión Rodríguez & Gómez Guinovart 2001 = José L. Aguirre Moreno, Nuria Andión Rodríguez, Xavier Gómez Guinovart 2001. Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega. *Procesamiento del lenguaje natural* 27, 13-20. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/download/3338/1826.

Alba-Salas 2012 = Josep Alba-Salas 2012. Colocaciones incoativas con *tomar* y *prender* en diacronía. *Revista de historia de la lengua española* 7, 3-38. https://doi.org/10.54166/rhle.2012.07.01.

Biblia medieval = Andrés Enrique-Arias (ed.) 2008-. *Biblia medieval*. http://bibliamedieval.es.

Brants 1996 = Thorsten Brants 1996. *TnT. A statistical part-of-speech tagger*. Universität des Saarlandes.

Brinton 2012 = Laurel J. Brinton 2012. Historical pragmatics and corpus linguistics: problems and strategies. Merja Kytö (ed.). *English corpus linguistics: Crossing paths*. Rodopi, 101-132.

Brown & Gilman 1960 = Roger Brown, Albert Gilman 1960. The pronouns of power and solidarity. Thomas A. Sebeok (ed.). *Style in language*. The MIT Press, 253-276.

Camus Bergareche 2006 = Bruno Camus Bergareche 2006. La expresión de la negación. Concepción Company Company (ed.). *Sintaxis histórica de la lengua española. Parte 1. La frase verbal. Vol. 2*. Universidad Nacional Autónoma de México, Fondo de Cultura Económica, 1163-1249.

CARRIZO = María C. Casado Lobato (ed.) 1983. *Colección diplomática del Monasterio de Carrizo*, I. Centro de Estudios e Investigación *San Isidoro*.

CE = Mark Davies (ed.) 2001-. *Corpus del español*. https://www.corpusdelespanol.org.

CHARTA = Pedro Sánchez-Prieto Borja (ed.) 2011-. *Corpus hispánico y americano en la red*. https://www.redcharta.es.

CODEA+ 2022 = Pedro Sánchez-Prieto Borja (ed.) 2022. *CODEA+ 2022. Corpus de documentos españoles anteriores a 1900*. https://www.corpuscodea.es.

Company Company 2006 = Concepción Company Company (ed.) 2006. *Sintaxis histórica de la lengua española. Parte 1. La frase verbal*. Universidad Nacional Autónoma de México, Fondo de Cultura Económica.

Company Company 2009 = Concepción Company Company (ed.) 2009. *Sintaxis histórica de la lengua española. Parte 2. La frase nominal*. Universidad Nacional Autónoma de México, Fondo de Cultura Económica.

Company Company 2014 = Concepción Company Company (ed.) 2014. *Sintaxis histórica de la lengua española. Parte 3. Adverbios, preposiciones y conjunciones. Relaciones interoracionales*. Universidad Nacional Autónoma de México, Fondo de Cultura Económica.

CORDE = Real Academia Española (ed.) 2008. *Corpus diacrónico del español*. http://corpus.rae.es/cordenet.html.

CORDEREGRA = Miguel Calderón Campos 2015. *El español del Reino de Granada en sus documentos (1492-1833). Oralidad y escritura*. Lang.

CORDIAM = Concepción Company Company, Virginia Bertolotti (eds.) 2016-. *Corpus diacrónico y diatópico del español de América*. https://www.cordiam.org/.

CORHEN = María J. Torrens Álvarez (ed.) 2019-. *Corpus histórico del español norteño*. https://corhen.es/.

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

DICCA XV = Coloma Lleal Galceran (ed.) 2013-. *Diccionari del castellà del segle XV a la Corona d'Aragó*. http://ghcl.ub.edu/diccaxv/.

DITECA = Inés Carrasco Cantos, Pilar Carrasco Cantos (eds.) 2011. *Diccionario de textos concejiles de Andalucía*. http://www.arinta.uma.es.

Domínguez Ortiz 2012 [1973] = Antonio Domínguez Ortiz 2012 [1973]. *Las clases privilegiadas en el Antiguo Régimen*. Akal.

Enrique-Arias 2012 = Andrés Enrique-Arias 2012. Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad. *Scriptum digital* 1, 85-106. https://doi.org/10.5565/rev/scriptum.37.

Fernández Reyes, Leyva Pérez & Lau Fernández 2011 = Francis de la C. Fernández Reyes, Exiquio C. Leyva Pérez, Rogelio Lau Fernández 2011. Consideraciones de diseño para una herramienta de análisis semántico. *Revista de lingüística teórica y aplicada* 49.1, 51-68. https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-48832011000100004.

Fernández-Ordóñez 2006 = Inés Fernández-Ordóñez 2006. La historiografía medieval como fuente de datos lingüísticos. Tradiciones consolidadas y rupturas necesarias. José J. de Bustos Tovar, José L. Girón Alconchel (eds.). *Actas del VI congreso internacional de historia de la lengua española. Madrid. 29 de septiembre - 3 de octubre de 2003. Vol. 2*. Arco, 1779-1807.

Fernández-Ordóñez 2009 = Inés Fernández-Ordóñez 2009. Orden de palabras, tópicos y focos en la prosa alfonsí. *Alcanate. Revista de estudios alfonsíes* 6, 139-172. https://dialnet.unirioja.es/servlet/articulo?codigo=2990478.

FreeLing = Lluís Padró (ed.) 2005-. *FreeLing*. https://nlp.lsi.upc.edu/freeling/.

Garachana 2014 = Mar Garachana 2014. Coordinación copulativa *e(t)/y* y disyuntiva *o*. Concepción Company Company (ed.). *Sintaxis histórica de la lengua española. Parte 3. Adverbios, preposiciones y conjunciones. Relaciones interoracionales. Vol. 2*. Universidad Nacional Autónoma de México, Fondo de Cultura Económica, 2337-2520.

García Moreno & Pueyo Mena 2017 = Aitor García Moreno, Francisco J. Pueyo Mena 2017. Etiquetado gramatical y lematización en el *Corpus histórico judeoespañol* (*CORHIJE*): Problemas, soluciones y resoluciones. *Scriptum digital* 6, 69-82. https://doi.org/10.5565/rev/scriptum.78.

García-Miguel 2022 = José M. García-Miguel 2022. Lingüística de corpus: de los datos textuales a la teoría lingüística. *Estudios de lingüística del español* 45, 11-42. https://doi.org/10.36950/elies.2022.45.8848.

Goded Rambaud & Ibáñez Moreno 2012 = Margarita Goded Rambaud, Ana Ibáñez Moreno 2012. El desarrollo de un etiquetado semántico-cognitivo para el procesamiento de estructuras léxicas de un corpus de notas de cata de vino. María J. Salinero Cascante, Elena González Fandos (eds.). *Vino y alimentación. Estudios humanísticos y científicos*. Universidad de La Rioja, 181-196.

Graña Gil 2000 = Jorge Graña Gil 2000. *Técnicas de análisis sintáctico robusto para la etiquetación del lenguaje natural*. Tesis doctoral, Universidad de La Coruña. https://ruc.udc.es/dspace/handle/2183/12358.

Gumperz 1982 = John J. Gumperz 1982. *Discourse strategies*. Cambridge University Press.

HESPERIA = Joaquín Gorrochategui (ed.) 1997-. *Banco de datos de lenguas paleohispánicas*. http://hesperia.ucm.es/.

Isasi Martínez et al. 2020 = Carmen Isasi Martínez, Leyre Martín Aizpuru, Santiago Pérez Isasi, Elena Pierazzo, Paul Spence 2020. *Edición digital de documentos antiguos: marcación XML-TEI basada en el criterio CHARTA*. Universidad de Sevilla.

Kabatek 2005 = Johannes Kabatek 2005. Tradiciones discursivas y cambio lingüístico. *Lexis* 29.2, 151-177. https://doi.org/10.18800/lexis.200502.001.

Koch & Oesterreicher 1985 = Peter Koch, Wulf Oesterreicher 1985. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36, 15-43.

Labov 1972 = William Labov 1972. *Sociolinguistic patterns*. University of Pennsylvania Press.

Llisterri Boix & Torruella Casañas 1999 = Joaquim Llisterri Boix, Joan Torruella Casañas 1999. Diseño de corpus textuales y orales. José M. Blecua, Gloria Clavería, Carlos Sánchez, Joan Torruella (eds.). *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Universidad Autónoma de Barcelona, 45-77.

LYNEAL = Hiroto Ueda. *LYNEAL. Letras y números en análisis lingüísticos*. https://h-ueda.sakura.ne.jp/lyneal/.

Mateu Ibars & Mateu Ibars 1980-1991 = Josefina Mateu Ibars, M. Dolores Mateu Ibars (eds.) 1980-1991. *Colectánea paleográfica de la Corona de Aragón. Siglos IX-XVIII*. Universidad de Barcelona.

McEnery, Xiao & Tono 2006 = Tony McEnery, Richard Xiao, Yukio Tono 2006. *Corpus-based language studies. An advanced research book*. Routledge.

Menéndez Pidal 1976 [1944] = Ramón Menéndez Pidal (ed.) 1976 [1944]. *Cantar de mio Cid. Texto, gramática y vocabulario*. 5a edición. Espasa-Calpe.

Montaner 1993 = Alberto Montaner (ed.) 1993. *Cantar de mio Cid*. 2a edición. Crítica.

Montoyo 2002 = Andrés Montoyo Guijarro 2002. *Desambiguación léxica mediante marcas de especificidad*. Tesis doctoral, Universidad de Alicante. http://rua.ua.es/dspace/handle/10045/3745.

Moreno 2002 = María Cristobalina Moreno 2002. The address system in the Spanish of the Golden Age. *Journal of Pragmatics* 34.1, 15-47.

Nica 2004 = Iulia Nica 2004. *El conocimiento lingüístico en la desambiguación semántica automática*. Tesis doctoral, Universidad de Barcelona.

ODE = Miguel Calderón Campos, María T. García Godoy 2019-. *Oralia diacrónica del español*. http://corpora.ugr.es/ode.

Oliver González 2021 = Antoni Oliver González. *Programación en Python para filólogos, lingüistas y traductores. 5. Etiquetado morfosintáctico*. Universitat oberta de Catalunya. https://xwiki.recursos.uoc.edu/wiki/matm21564es/view/5.%20Etiquetado%20morfosintático/.

Ribera, Molina & Pla 2000 = Xavier Ribera, Antonio Molina, Ferrán Pla 2000. Herramienta para el etiquetado léxico y análisis sintáctico de textos orientado a la construcción de corpus supervisados. *Procesamiento del lenguaje natural* 26, 119-124. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/3437/1925.

Rodríguez Molina & Octavio de Toledo y Huerta 2017 = Javier Rodríguez Molina, Álvaro Octavio de Toledo y Huerta 2017. La imprescindible distinción entre texto y testimonio: el *CORDE* y la fiabilidad lingüística. *Scriptum digital* 6, 5-68. https://doi.org/10.5565/rev/scriptum.73.

Sánchez Sánchez & Domínguez Cintas 2007 = Mercedes Sánchez Sánchez, Carlos Domínguez Cintas 2007. El banco de datos de la Real Academia Española: *CREA* y *CORDE*. *Per Abbat. Boletín filológico de actualización académica y didáctica* 2, 137-148. https://dialnet.unirioja.es/servlet/articulo?codigo=2210249.

Sanjurjo González 2017 = Hugo Sanjurjo González 2017. *Creación de un framework para el tratamiento de corpus lingüísticos*. Tesis doctoral, Universidad de León. https://buleria.unileon.es/handle/10612/6920.

Sierra Martínez 2017 = Gerardo E. Sierra Martínez 2017. *Introducción a los corpus lingüísticos*. Universidad Nacional Autónoma de México.

SLI-Tagger = Seminario de lingüística informática, Grupo TALG 2016-2022. *SLI-Tagger - Anotación de textos en galego, portugués, catalán, español e inglés*. https://ilg.usc.gal/tagger/.

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815

Suárez Cueto 2004 = Armando Suárez Cueto 2004. *Resolución automática de la ambigüedad semántica de las palabras mediante modelos de probabilidad de máxima entropía*. Tesis doctoral, Universidad de Alicante. http://rua.ua.es/dspace/handle/10045/4070.

Taulé Delor et al. 2006 = Mariona Taulé Delor, Joan Castellví Vives, M. Antonia Martí Antonín, Juan Aparicio 2006. Fundamentos teóricos y metodológicos para el etiquetado semántico de CESS-CAT y CESS-ESP. *Procesamiento del lenguaje natural* 37, 75-82. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2735/1253.

Thera = Centre de Llenguatge i Computacio (CLiC) (ed.) 2010. *Thera*. https://clic.ub.edu/.

Tognini-Bonelli 2001 = Elena Tognini-Bonelli 2001. *Corpus linguistics at work*. Benjamins.

TreeTagger = Helmut Schmid 1994-. *TreeTagger – a part-of-speech tagger for many languages*. https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

Vilar Bohigues 2010 = Laura Vilar Bohigues 2010. *Proceso de etiquetado de un corpus digital*. Proyecto final de carrera, Universidad Politécnica de Valencia. https://riunet.upv.es/handle/10251/10298.

WordNet = Princeton University 1995-. *WordNet. A lexical database for English*. https://wordnet.princeton.edu.

Wright 1989 [1982] = Roger Wright 1989 [1982]. *Latín tardío y romance temprano en España y la Francia carolingia*. Gredos [traducción de *Late Latin and early romance in Spain and Carolingian France*. Cairns].

Zieliński 2021 = Andrzej Zieliński 2021. La fórmula de tratamiento con *señor*, ¿posible germanismo? *Romanica Cracoviensia* 21.1, 33-42. https://doi.org/10.4467/20843917RC.21.003.13671.

Zieliński, en prensa= Andrzej Zieliński, en prensa. De 'buscar' a 'querer'. Notas sobre el origen del verbo *querer* en las lenguas iberorromances. Marzena Chrobak, Anna Wolny (eds.). *Żegluając po świecie romańskim. Studia filologiczne*. Wydawnictwo Uniwersytetu Jagiellońskiego.

Zieliński & Espinosa Elorza, en prensa = Andrzej Zieliński, Rosa M. Espinosa Elorza, en prensa. Estructuras semiperifrásticas y perifrásticas con verbos modales. Concepción Company Company (ed.). *Sintaxis histórica de la lengua española. Parte 4. Estructura argumental, estructura informativa y discurso. Tradiciones discursivas y géneros textuales*. Universidad Nacional Autónoma de México, Fondo de Cultura Económica.

Zumthor 2001 [1987]= Paul Zumthor 2001 [1987]. *A letra e a voz: a literatura medieval*. Companhia das Letras [traducción de *La lettre et la voix. De la "littérature" médiévale*. Seuil].

Xiga = Seminario de lingüística informática, Grupo TALG 2017. *Corpus XIGA de textos de informática e telecomunicacións en galego*. http://ilg.usc.gal/ctg/corpus.php.

DOI: https://doi.org/10.25364/19.2024.12.4 ISSN: 2663-9815