Ciencias Sociales y Humanidades Digitales Aplicadas

Casos de estudio y perspectivas críticas

Esteban Romero Frías
Lidia Bocanegra Barbecho
(eds.)

Know metrics

Ciencias Sociales y Humanidades Digitales Aplicadas

Casos de estudio y perspectivas críticas

Esteban Romero Frías
Lidia Bocanegra Barbecho
(eds.)

Ciencias Sociales y Humanidades Digitales Aplicadas

Casos de estudio y perspectivas críticas

Esteban Romero Frías

Lidia Bocanegra Barbecho

(eds.)

Medialab UGR, Universidad de Granada

Septiembre 2018

eug EDITORIAL UNIVERSIDAD DE GRANADA

DOWN HILL publishing

CIENCIAS SOCIALES Y HUMANIDADES DIGITALES APLICADAS: CASOS DE ESTUDIO Y PERSPECTIVAS CRÍTICAS
APPLIED DIGITAL HUMANITIES AND SOCIAL SCIENCE: CASES STUDIES AND CRITICAL PERSPECTIVES

Este trabajo ha sido posible gracias a la financiación del proyecto "Knowmetrics: evaluación del conocimiento en la sociedad digital" dentro de las ayudas de la Fundación BBVA a equipos de investigación científica 2016.

COORDINAN:
Esteban Romero Frías
Lidia Bocanegra Barbecho

EDITAN:
Universidad de Granada: ISBN 978-84-338-6318-8
Downhill Publishing (NY): ISBN-13: 978-0-9897361-7-6

PATROCINAN:
Fundación BBVA - Proyecto Knowmetrics
Medialab UGR- Laboratorio de Investigación en Cultura y Sociedad Digital

DISEÑO Y MAQUETACIÓN:
Javier Cantón
Tipografías abiertas utilizadas en la confección de este libro:
LLPixel de Markus Schröppel
EB Garamond de Georg Duffner
Fecha de publicación: 30 de septiembre de 2018

Índice

Introducción
Esteban Romero Frías y Lidia Bocanegra Barbecho..... 12

Los entornos de aprendizaje personal y el alumnado universitario
Mirian Hervás Torres..... 14

Tecnologías de control y vigilancia en el Arte Contemporáneo
José Luis Lozano Jiménez..... 40

Autocreación de *video abstracts* como parte de la investigación multimodal
Paloma Marín Arraiza y Attila Dávid Molnár 66

The Non Una di Meno Feminist Movement in Italy: Connective or Collective?
Tommaso Trillò..... 85

La sexualidad humana en la era digital
Anes Ortigosa Blázquez y María José de la Torre Ávalos ... 110

El dibujo, uno de los pilares del mundo de los videojuegos
Andrés Domenech Alcaide..... 130

Islam 2.0: Mapas conceptuales para un Mediterráneo en mutación
Luz Gómez, Elena Arigita, Laura Galián y Jesús Zanón 151

Criterios para la evaluación de la implicación del público en la ciencia a través de la Web 2.0
Lourdes López-Pérez y María Dolores Olvera-Lobo 175

Ciudadanía digital responsable frente al derecho al olvido
Marta Grande Sanz 202

La educación virtual para el estudio de las Ciencias Jurídicas: un MOOC para el aprendizaje de la Historia del Derecho y de las Instituciones
Marina Rojo Gallego-Burín y Araceli María Rojo Gallego-Burín.....223

Ciberbullying y territorios digitales: claves para un abordaje desde la perspectiva educativa
Luis Miguel Rondón García y Juan Lorenzo Bermúdez Díaz..... 245

La Barranquilla de Voces: ejercicios de una re-cartografía espacial
Luis Alfonso Barragán Varela 267

Turismo y nuevas tecnologías, una pareja ¿bien avenida?
Inmaculada Mengual Bernal 284

¿Una historia del futuro? El debate de los economistas sobre la cuarta revolución industrial
Simone Fari..... 300

Diseño y lanzamiento de un MOOC como instrumento formativo digital gratuito para inclusión de migrantes y refugiados. Análisis preliminar de resultados
Bianca Vitalaru, Carmen Valero-Garcés y Raquel Lázaro Gutiérrez..... 325

El *Archivo Digital Valle-Inclán*: aplicación web
Carmen E. Vílchez Ruiz 361

Análisis de cursos abiertos, masivos y en línea (MOOC) como ecologías de formación. Potencialidades del estudio de caso como metodología
Ramón Montes Rodríguez..... 386

Sociedad digital, identidad y diáspora: el caso del Museo Palestino
Javier Guirado Alonso 410

La digitalización de la Estadística General del Reino, 1817-1820: un proyecto en construcción
Miguel Á. Bringas, Íñigo del Mazo y Guillermo Mercapide 428

El proyecto Artapp: diseño e implementación de una plataforma para la promoción, difusión y legitimación del artista
César González Martín, Ana García López y Belén Mazuecos Sánchez..... 457

Los datos de la ficción: visualización de las relaciones discursivas entre personajes en *Crimen y Castigo*
Benamí Barros García 489

Investigando Granada a través de Instagram
Fco. Javier Cantón Correa y Jordi Alberich Pascual..... 507

Where the researcher cannot get: open platforms to collaborate with citizens on cultural heritage research data
Maurizio Toscano 538

Las Humanidades Digitales y los corpus diacrónicos en línea del español: problemas y sugerencias
Rocío Díaz Bravo 562

***Altmetric beauties:* ¿cuáles son los trabajos científicos con mayor impacto en las redes sociales?**
Wenceslao Arroyo Machado y Daniel Torres-Salinas ... 587

Díaz Bravo, R. (2018). Las Humanidades Digitales y los corpus diacrónicos en línea del español: problemas y sugerencias. En E. Romero Frías y L. Bocanegra Barbecho (Eds.), *Ciencias Sociales y Humanidades Digitales Aplicadas* (pp. 562-586). Granada, España: Universidad de Granada [ISBN: 978-84-338-6318-8]. New York, USA: Downhill Publishing [ISBN: 978-0-9897361-7-6].

Las Humanidades Digitales y los corpus diacrónicos en línea del español: problemas y sugerencias

Rocío Díaz Bravo
Universidad de Granada
rociodiazbravo@ugr.es

Resumen

El objetivo de este capítulo es exponer los resultados de un análisis de corpus diacrónicos del español, centrado especialmente en lematizadores y etiquetadores morfosintácticos, teniendo en cuenta las necesidades y perfiles de sus usuarios.

Los resultados de mis entrevistas con investigadores que trabajan en diferentes proyectos y líneas de investigación de la Historia de la Lengua Española ponen de manifiesto la necesidad de interfaces intuitivas, de corpus lematizados y etiquetados, con opciones

Este trabajo se enmarca en el proyecto de referencia FFI2017-83400-P (MINECO/AEI/FEDER, UE).

de búsquedas que permitan todo tipo de estudios lingüísticos (en todos los niveles, incluida la variación sociolingüística).

En los últimos años se ha incrementado el número de recursos digitales y de corpus diacrónicos del español. A pesar de los avances de los grandes corpus diacrónicos del español –Corpus del Español de Mark Davies (CdE) y Corpus del Nuevo Diccionario Histórico de la Real Academia Española (CDH)–, todavía presentan problemas desde el punto de vista textual y tecnológico.

A través del caso práctico de *vos* en la historia del español, pretendo demostrar que los grandes corpus del español no son apropiados para muchos tipos de Investigación Lingüística y que, además, es necesario revisar los datos manualmente.

Después de analizar lematizadores y etiquetadores morfosintácticos de español anterior al siglo XX, y de probar su precisión aplicándolos a textos de diferentes periodos, se puede concluir que también deben ser mejorados. Asimismo, el único lematizador de español anterior al siglo XX que sigue estándares internacionales, Freeling, no ofrece una interfaz amigable.

Entre las soluciones sugeridas, debe subrayarse la importancia de estándares internacionales (como TEI para la edición de textos y para los metadatos, o EAGLES para la etiquetación morfosintáctica) por razones de transferibilidad y preservación, así como la necesidad de aumentar la colaboración y el entendimiento entre disciplinas (Humanidades Digitales, Lingüística Computacional e Historia de la Lengua Española).

Palabras clave: Humanidades Digitales, Lematización, Etiquetación Morfosintáctica, Corpus Diacrónicos, Historia de la Lengua Española

Abstract

The aim of this paper is to show the results of an analysis of Spanish diachronic online corpora, with a particular emphasis on

lemmatisers and PoS (part of speech) taggers, taking into account users' needs and backgrounds.

The results of my interviews with scholars working on different projects and areas within the history of Spanish have shown the need for intuitive user-friendly interfaces, lemmatised and annotated corpora, as well as advanced search options that allow different types of linguistic research (at all linguistic levels, including sociolinguistic variation).

In recent years there has been an increasing number of digital resources and diachronic corpora of Spanish. Despite the advances of very large Spanish diachronic corpora –Corpus del Español by Mark Davies (CdE) and Corpus del Nuevo Diccionario Histórico by Real Academia Española (CDH)–, from a textual and a technological point of view, they still exhibit problems. Through the case study of *vos* in the History of Spanish, I aim to demonstrate that large diachronic corpora of Spanish are not suitable for many types of linguistic research and in addition the data need to be revised manually. After analysing lemmatisers and PoS taggers for pre-20th century Spanish and testing them with texts from different periods in the History of Spanish, it can be concluded that they also need to be improved in terms of accuracy. Furthermore, the only lemmatiser of pre-20th century Spanish that follows international standards, Freeling, is not user friendly.

Among the solutions suggested, it is relevant to highlight the importance of international standards (such as TEI for editing texts and metadata, or EAGLES for PoS tagging) for reasons of transferability and preservation; as well as the need for greater collaboration and understanding between disciplines (Digital Humanities, Computational Linguistics and History of the Spanish Language).

Keywords: Digital Humanities, Lemmatisation, Part of Speech Tagging, Diachronic Corpora, History of the Spanish Language

Introducción, objetivos y metodología

La era digital está afectando a la investigación en todos los campos del saber, incluidas las Humanidades. En el caso de la Filología Hispánica y de la Historia de la Lengua Española, cada vez son más numerosos los proyectos de investigación que se están beneficiando de los métodos, recursos y herramientas de las Humanidades Digitales. A pesar del número creciente de recursos digitales de investigación en la Historia de la Lengua Española, la bibliografía sobre sus aspectos tecnológicos es escasa, así como el análisis de las necesidades de los investigadores y usuarios de dichos recursos. Por otro lado, estos recursos electrónicos no aprovechan todo el potencial que ofrecen las nuevas tecnologías. Por todo ello, es importante seguir fomentando el crecimiento de las Humanidades Digitales en la Historia de la Lengua Española, sobre todo, en la creación de recursos y herramientas que faciliten la investigación y la docencia de sus distintas áreas de especialización y niveles lingüísticos y que permitan nuevos caminos de exploración científica.

Los objetivos de esta investigación son, principalmente, demostrar que los corpus lingüísticos deben mejorar considerablemente en aspectos textuales y digitales; identificar opciones y funcionalidades que se podrían incorporar a los corpus para facilitar la investigación lingüística, teniendo en cuenta las necesidades de los usuarios; evaluar la fiabilidad de los lematizadores y etiquetadores del español anterior al siglo XX; proponer mejoras.

Para realizar esta investigación he empleado métodos mixtos (un estudio cualitativo y cuantitativo). Por una parte, he entrevistado a dieciocho investigadores[1] que trabajan en diferentes proyectos y líneas de investigación de la Historia de la Lengua Española, con objeto de analizar sus necesidades de investigación y prácticas habituales con respecto al uso de nuevas tecnologías. Además de

[1] Entre ellos, uno de los investigadores de los corpus de la RAE, de cuya entrevista se presentan algunos datos en este trabajo.

realizar preguntas específicas, observé cómo usan diversos recursos digitales y herramientas computacionales. Para ello, mostré recursos digitales y herramientas computacionales muy diferentes entre sí como ejemplos de diversas funcionalidades e interfaces. Entre los resultados más relevantes del análisis se pueden señalar la necesidad de interfaces intuitivas y amigables, de opciones de búsqueda avanzada, de corpus etiquetados desde el punto de vista gramatical (para los estudios de morfosintaxis); y el resultado más relevante, compartido por todos los investigadores entrevistados, es la necesidad de corpus lematizados para todo tipo de investigación lingüística (Díaz-Bravo, 2015: 379-381). Teniendo en cuenta las necesidades de los investigadores, he analizado los principales corpus del español, así como los lematizadores y etiquetadores morfosintácticos del español anterior al siglo XX.

Características de los grandes corpus diacrónicos del español: CORDE, CdE, CDH

Como los corpus son conjuntos extensos de textos escritos, han tenido un impacto muy significativo en la lingüística histórica, ya que permiten la investigación de enormes cantidades de datos de manera automática. En el caso del español, existen tres grandes corpus que intentan cubrir todos los períodos de esta lengua:

1. El Corpus del Español de Mark Davies (histórico-géneros), de 100 millones de palabras procedentes de los siglos XIII-XX, es especialmente deficiente en su composición textual. Por ejemplo, según aparece especificado en el archivo que contiene la información de las fuentes que componen el corpus, se incluyen las obras completas de Gonzalo de Berceo. Sin embargo, no aparecen los datos bibliográficos de dichas obras, sino únicamente un enlace –que no funciona al menos desde 2011– al sitio web

Geocities. Por otra parte, aunque permite comparar géneros (académico, periodístico, ficción y oral), normalmente no es posible comparar diversos géneros a través de la historia (por ejemplo, los textos académicos del siglo XX se limitan a la enciclopedia *Encarta*). No es posible realizar búsquedas por países o zonas en este corpus, pero sí en el Corpus del Español de web-dialectos (compuesto por textos de los siglos XX-XXI, por lo que no es útil para el estudio de la historia de la lengua).

2. La Real Academia Española ha creado dos corpus para el estudio de la historia de la lengua:
    - Primero, el CORDE, de 250 millones de palabras, usado por numerosos investigadores, que es especialmente deficiente en aspectos tecnológicos.
    - Más reciente y más avanzado tecnológicamente, aunque conserva algunas de las deficiencias del CORDE, es el CDH.

En la siguiente tabla se puede observar un resumen de las principales características de cada uno de estos corpus:

	Textos	Búsquedas	Tecnología	Concordancias	Gráficos
CORDE (RAE, finales de los 90)	250 millones de palabras 1200-1975 Obras transcritas con muy diversos criterios	Género, temas, fecha (pestañas desplegables) Países	SGML (antecesor de XML) Servicio de indexación de Microsoft	KWIC (PCEC) No lematización	NO
CdE (Davies, 2001)	100 millones de palabras 1200-2000 ¿Fiabilidad de los textos? ¿Representatividad?	Género, fecha, lemas	Base de datos relacional	KWIC (PCEC) Lematización, categoría gramatical	Barra Frecuencia relativa Distribución por siglos y por géneros
CDH (RAE, 2012)	Corpus nuclear: 62 millones de palabras (España: 38 millones América: 24 millones)	Zonas, países, fecha, lemas	XML	KWIC (PCEC) Lematización, categoría gramatical	Circulares Frecuencia absoluta Distribución por periodo, zona, país

Tabla 1: Resumen de las características de los grandes corpus diacrónicos del español. Fuente: elaboración propia.

Todos estos corpus han sido muy ambiciosos en cuanto a la cantidad de palabras, pero no en la calidad (tanto desde el punto de vista textual como digital). Como consecuencia, a pesar de que son recursos enormemente valiosos para muchos investigadores, deben mejorar significativamente, ya que pueden llevar a conclusiones erróneas, como quisiera demostrar a través del caso práctico del estudio de *vos* en la historia del español, origen del voseo latinoamericano.

Caso práctico: estudio del pronombre *vos* con los corpus diacrónicos del español

Los estudios de lingüística histórica pueden ser de diversa índole y se caracterizan por su fuerte estructuración: por niveles lingüísticos (ortográfico y fonético-fonológico, morfosintáctico, léxico, pragmático-discursivo), o estudios de cambio lingüístico y variación (diacrónicos, diatópicos o geográficos, diastráticos o sociales, diafásicos o contextuales).

Una posible investigación desde el punto de vista diacrónico sería estudiar la evolución de las formas de tratamiento, en

concreto, la evolución del pronombre *vos*, respondiendo a las preguntas de investigación de cómo y cuándo llegó a desvalorizarse en España, hasta el punto de que desapareció, frente a otros países. Como sabemos, la situación fue muy diferente en América Latina, donde *vos* está incluso aceptado en la norma culta en los países que tenían menos contacto con España durante la colonización, en los que actualmente se habla la variedad austral (Argentina, Uruguay, Paraguay). La siguiente tabla nos muestra un resumen simplificado de la evolución de *vos*:

2ª persona		**Latín clásico**	**Español medieval**	**Siglos XV-XVI**	**Español moderno**
Singular	**Familiar**	TÚ	TÚ	TÚ VOS	TÚ VOS (Hispanoamérica)
	Respeto		VOS	VOS VUESTRA MERCED	 USTED
Plural	**Familiar**	VOS	VOS	VOSOTROS	VOSOTROS (España)
	Respeto			VUESTRAS MERCEDES	USTEDES

Tabla 2: Evolución de vos *en la historia del español. Fuente: Elaboración propia.*

Otra posibilidad es realizar un estudio diatópico, es decir, un análisis que nos permita conocer en qué países o zonas se vosea y en cuáles no. Puesto que la variación lingüística no ocurre de manera aislada, sino en cadena (Koch y Oesterreicher, 2007: 39), sería relevante combinar los diferentes tipos de variación lingüística; por ejemplo: variación diatópica y diacrónica (países y fechas); variación diatópica y diastrática (origen geográfico y social de los hablantes, respectivamente); y todo ello se podría combinar con el último escalón de la cadena variacional, la variación diafásica, que tiene en cuenta si el registro o el contexto de la comunicación es más o menos formal. Podríamos plantear preguntas de investigación como las siguientes: ¿desde-hasta qué fecha se ha usado *vos* en los diferentes países hispanohablantes; qué condición social

poseen los hablantes voseantes en cada uno de esos países o zonas; en qué registros o contextos se emplea esta forma de tratamiento?

Con los corpus diacrónicos del español, sería imposible una investigación en la que se estudie la variación diastrática y diafásica de una novela en la que su autor intenta caracterizar a sus personajes por su forma de hablar, teniendo en cuenta su condición social y las distintas situaciones comunicativas (desde más formales hasta más íntimas). Se debe principalmente a que la obra aparece etiquetada en su totalidad como "ficción", sin distinguir los diferentes fragmentos en los que existe una imitación de la oralidad, como los diálogos o el estilo directo; asimismo, en ningún caso se incluye la condición social de los autores de los textos en los corpus, y mucho menos, información sobre los personajes que aparecen en los mismos, cuya codificación sería una tarea de gran complejidad. A modo de ejemplo, resumo los resultados de mi investigación de las formas de tratamiento en el *Retrato de la Loçana andaluza* de Francisco Delicado (1530?) (Díaz-Bravo, 2010: 372-381):

- Con respecto a la variación diastrática, *vuestra merced* se usa para dirigirse a un interlocutor superior, *vos* es un tratamiento polisémico que se acerca tanto a los valores de *vuestra merced* como a los de *vos*, *tú* se usa para dirigirse a un interlocutor de condición social inferior. Se encuentran ejemplos de uso sexista de la lengua en esta obra, ya que una madre se dirige a su hijo hablándole de *vos*, mientras que para su hija utiliza *tú*.
- En cuanto a la variación diafásica, la protagonisa –la prostituta Lozana–, en situaciones formales en las que está tratando el pago de sus servicios con sus clientes, se dirige a ellos con el tratamiento de *vuestra merced*; en cambio, en contextos de intimidad, se intercambian el tratamiento de *vos*. También es destacable el cambio de identidad de Rampín –y el consecuente cambio en las formas de tratamiento empleadas–: como novio de

> Lozana, en situaciones en las que actúan como tal, se intercambian el tratamiento *vos*; sin embargo, cuando él finge ser su criado delante de otros personajes, ella se dirige a Rampín con el tratamiento propio para hablarle a un "inferior": *tú*.

Otro aspecto importante que dificulta el estudio del tratamiento *vos* en los corpus diacrónicos del español es el hecho de que dicho pronombre normalmente no se usa explícitamente, sino que aparece implícito en las desinencias verbales (por ejemplo, en el imperativo, en fragmentos de la conversación entre Lozana y Tía del mamotreto II del *Retrato de la Loçana andaluza*: "mirá, señora tía"; "pensá, señora"); o en la concordancia con los posesivos (por ejemplo, en el mismo mamotreto, refiriéndose a "vuestro padre": "me dixeron que se casó por amores con vuestra madre") (Díaz-Bravo, en prensa). La única manera de poder realizar búsquedas que incluyan los posesivos asociados al voseo y, sobre todo, las desinencias verbales, es contar con un corpus lematizado y anotado morfosintácticamente, que además permita recuperar la información que ha sido etiquetada.

En definitiva, ¿nos permiten los grandes corpus diacrónicos del español realizar investigaciones como las que se acaban de mencionar? La respuesta es negativa; aunque, en algunos casos, sí es posible, pero con mucha investigación y revisión adicional.

Problemas que presentan los corpus diacrónicos del español

En este apartado pretendo demostrar que los corpus lingüísticos pueden conducir a conclusiones erróneas.

Desde una perspectiva diacrónica, podemos observar en los gráficos pertenecientes a tres recursos digitales que sus resultados son complemente diferentes entre sí cuando buscamos la

evolución de *vos* a través de la historia. Los siguientes gráficos no reflejan la realidad lingüística:

Figura 1: Gráfico que presenta los resultados de la evolución diacrónica de vos *y de su distribución por géneros en el CdE. Fuente: CdE*

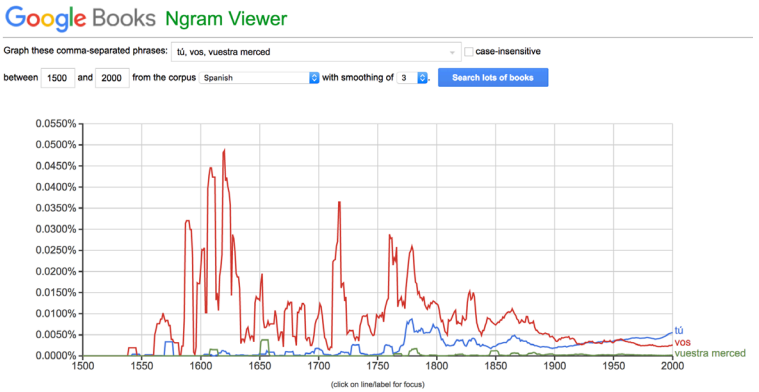


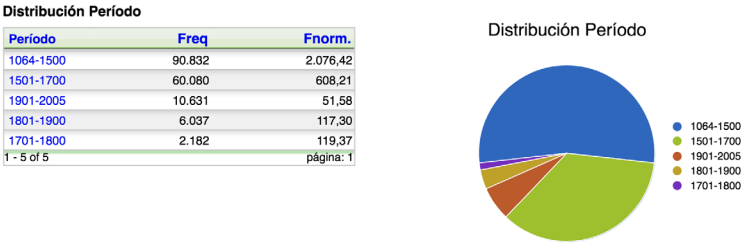
Figura 2: Gráfico que presenta los resultados de tú, vos *y* vuestra merced *a través de un eje cronológico, según su frecuencia en Google Books. Fuente: Ngram Viewer*

El primer gráfico, del CdE, muestra la frecuencia relativa de *vos* a través de una visualización cronológica muy eficiente. Sin embargo, la mayoría de los textos incluidos en este corpus son de España (y no de Hispanoamérica), lo que explica la escasez de ejemplos a partir del siglo XVII, fecha en la que en España se había convertido en un tabú impronunciable (Díaz-Bravo, 2007).

El segundo gráfico procede de Ngram Viewer (Google Books), que no ha sido diseñado para la investigación lingüística histórica y que no sirve para analizar el español medieval y clásico; por tanto, los resultados tampoco reflejan la realidad lingüística.

Sin embargo, es un ejemplo de una visualización eficiente de la evolución de variantes competidoras (*tú, vos* y *vuestra merced*), que debería incluirse en corpus diacrónicos, como señalaron los investigadores en mis entrevistas.

Mucho menos efectivo desde el punto de vista visual es el gráfico que muestra la variación diacrónica por siglos en el CDH, pues se muestran en forma circular y ni siquiera en orden cronológico:



Distribución Período

Período	Freq	Fnorm.
1064-1500	90.832	2.076,42
1501-1700	60.080	608,21
1901-2005	10.631	51,58
1801-1900	6.037	117,30
1701-1800	2.182	119,37
1 - 5 of 5		página: 1

Figura 3: Gráfico que presenta la distribución cronológica de vos en el CDH. Fuente: CDH

Otro motivo que contribuye a que los resultados no sean precisos es el hecho de que la lematización del corpus se ha realizado de manera semiautomática y, como se explica en el "Manual de consulta en línea", se ha decidido "primar la cobertura sobre la precisión" (RAE-CDH: 9). A medida que se vaya elaborando el *Nuevo Diccionario Histórico del Español*, para cuyo fin se creó el CDH, se irá revisando esta lematización y anotación automática. Mientras tanto, los resultados deben consultarse con precaución y necesitan una revisión adicional. Por ejemplo, entre los resultados de *vos* como pronombre personal, aparecen numerosos ejemplos en las concordancias que en realidad se refieren al sustantivo *voz* o incluso a la palabra latina *vox*: no han sido desambiguadas correctamente en el proceso automático de lematización y anotación morfosintáctica, ni han sido revisadas posteriormente por lingüistas o filólogos. En la siguiente imagen es posible visualizar las diversas variantes gráficas del lema *vos*: con *v,* con *u*, con *b* y

su categoría gramatical (pronombre personal); los datos son incorrectos, como percibe al analizar las concordancias en formato de PCEC (palabra clave en contexto).

Forma	Categoría	f	f Rel
vos	pronombre personal	32939	77.6
uos	pronombre personal	5018	11.82
Vos	pronombre personal	2206	5.19
vós	pronombre personal	1989	4.68
bos	pronombre personal	148	0.34
Vós	pronombre personal	97	0.22
Uos	pronombre personal	11	0.02
vox	pronombre personal	7	0.01
VOS	pronombre personal	5	0.01
Uós	pronombre personal	5	0.01

obras quél auíe fechas. E en	bos	de Jhesu Cristo su fijo, cor
Santo doquiere espiraua, e la	bos	dél oyen nin sabíen dónde
ıbrél fueron dichas; díxolas la	bos	del Padre, en que fizo ente
ecta dixo en otro lugar que la	bos	de Dios, del Sennor de la

Figura 4: Resultados de la variación gráfica del pronombre personal vos *y algunos ejemplos de PCEC para la forma* bos. *Fuente: CDH*

El único de los tres corpus que ofrece gráficos con información diatópica (por áreas dialectales y por países) es el CDH, como se observa a continuación:

Distribución Zona

Zona	Freq	Fnorm.
España	156.881	510,74
Río de la Plata	5.368	227,09
México y Centroamérica	3.724	127,11
Andina	1.166	66,12
Caribe continental	1.098	68,39
Chilena	1.061	98,93
Antillas	417	42,32
Filipinas	38	20,19
Estados Unidos	9	4,22

1 - 9 of 9 página: 1

Distribución País

País	Freq	Fnorm.
España	156.881	510,74
Argentina	4.619	233,98
México	2.186	103,74
Chile	1.061	98,93
Colombia	922	108,68
Perú	781	61,72
Guatemala	682	211,87
Uruguay	546	218,76
Costa Rica	356	205,39
Cuba	289	44,42

1 - 10 of 22 página: 1 2 3

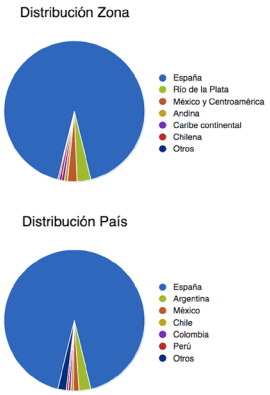


Figura 5: Gráfico que presenta la distribución diatópica de vos *en el CDH. Fuente: CDH*

Como se puede observar al contrastar los datos numéricos que aparecen en la tabla de la izquierda y los gráficos de la derecha,

estos últimos muestran la frecuencia absoluta en vez de la frecuencia normalizada (número de ocurrencias por millón de palabras), lo cual distorsiona los resultados. En términos absolutos, el número de ocurrencias totales de *vos* en España es mucho más elevado que en el resto de los países. Este resultado se explica porque más de la mitad de los textos del corpus son de España, por lo que el corpus está desequilibrado. Además, el gráfico debería reflejar los resultados de la frecuencia normalizada, ya que la frecuencia absoluta depende mucho del número de palabras incluidas para cada variable (países, períodos, etc.). Este problema se repite en todos los gráficos generados por el CDH, así como por el CORPES XXI (el nuevo corpus de la RAE, que ofrece la misma interfaz que el CDH, pero que contiene textos de español actual, y que además ofrece la distribución por temas y por tipologías).

Finalmente, es complejo o casi imposible estudiar la variación diastrática y diafásica con los corpus actuales, pues es necesario realizar mucho trabajo adicional. El CORDE ofrece una opción de búsqueda por tema, que en principio permitiría estudiar la variación diafásica (por registros) o la variación discursiva, pero es una clasificación muy extensa y difícil de usar (clasificación con numerosos subapartados que aparecen en una ventana de la que solo se visualizan las primeras líneas; sería de uso más fácil si se sustituyera por una pestaña desplegable). De cualquier forma, ninguno de los investigadores que han participado en mis encuestas usa dicha clasificación temática. Sorprendentemente, el sucesor del CORDE –el CDH– no ofrece ninguna opción de búsqueda o filtro que permita analizar la variación diastrática ni diafásica. Como se describió anteriormente, el CdE permite filtrar la información por "géneros" (gráfico 1), aunque la lista de géneros y periodos suele coincidir (por ejemplo, solo existe el género oral para textos del siglo XX; además, se trata fundamentalmente de corpus de habla culta).

En definitiva, ¿cuáles son los problemas de los corpus diacrónicos del español disponibles en línea en la actualidad? Desde el punto de vista filológico: los textos están editados con diversos criterios de edición, que no son explícitos, y algunos de los textos incluidos no siguen criterios filológicos, como han tenido que reconocer los propios coordinadores de los corpus. Además, existe un desequilibrio en la composición de textos representativos de diferentes siglos, géneros, áreas geográficas (la mayoría son de España), etc. Los metadatos deben ser más precisos. Desde el punto de vista tecnológico, me gustaría destacar los siguientes aspectos: la falta de análisis de las necesidades de los usuarios: el escaso aprovechamiento del potencial que ofrecen las nuevas tecnologías; la inexistencia de grandes corpus totalmente lematizados y anotados lingüísticamente posteriormente revisados por lingüistas y filólogos, por lo que es muy difícil especular las posibles variantes de una búsqueda. Las consecuencias de estos problemas son la falta de fiabilidad de los resultados, lo cual hace necesaria la revisión manual de los datos, y la escasa o nula idoneidad de los corpus para muchos tipos de investigación lingüística.

Sugerencias: propuestas de mejora para los corpus diacrónicos del español

Como he demostrado, los corpus diacrónicos de español deben mejorar para facilitar la investigación de la historia de la lengua española y de sociolingüística histórica. No sirven para el estudio de la variación diatópica o geográfica, ya que los metadatos incluidos en los corpus que permiten filtrar los resultados por países (CORDE y CDH) atienden al lugar de publicación de la obra y no al lugar de nacimiento del autor. Por tanto, no refleja la variedad del autor de la obra, por lo que no se tienen en cuenta las necesidades de los investigadores interesados en estudios dialectales. Tampoco son apropiados para el estudio de la variación diastrática

ni diafásica. Los metadatos deberían ser más precisos y deberían incluir detalles biográficos de los autores que nos permitan conocer su origen social, además de una tipología textual como, por ejemplo, la ofrecida en el Corpus Informatizat del Català Antic (permite buscar por "estilo directo", lo cual es muy útil para estudios de oralidad).

En todos los casos, es indispensable que existan metadatos precisos y detallados, que se puedan recuperar a través de una interfaz con opciones de búsqueda que permitan un análisis multivariable, lo cual sería ideal para el estudio de la variación y el cambio lingüístico. Este problema se solucionaría si los textos estuvieran marcados con el estándar internacional TEI (Text Encoding Initiative), basado en el estándar XML (eXtensible Markup Language). XML es una lengua de marcación que sirve para almacenar datos altamente estructurados, que se pueden intercambiar entre diferentes plataformas. La información que ha sido debidamente etiquetada puede luego ser recuperada en búsquedas específicas. Por ejemplo, si se etiquetan metadatos como la fecha y lugar de publicación del texto, autor (oficio, lugar de origen, en caso de que existan datos suficientes), será posible crear índices y realizar búsquedas de datos concretos. Para la codificación de los textos usando las etiquetas TEI, se puede usar el editor de XML Oxygen, que gracias a su potente sistema de validación permite evitar y corregir fácilmente errores de estructura en las etiquetas. El estándar internacional TEI ha sido especialmente diseñado para las Humanidades y contiene directrices específicas tanto para la edición digital de textos como para los corpus lingüísticos.

La siguiente tabla muestra un resumen de los problemas y las soluciones sugeridas desde una perspectiva tecnológica para posibilitar y mejorar los estudios de variación sociolingüística en la historia de la lengua española:

Tipos de variación lingüística	Análisis con corpus	Soluciones (tecnología)
¿? **Variación diacrónica** ✔ **1ª documentación**	Resultados incorrectos en la distribución por periodos Criterio cronológico: el más usado	Lematización revisada por lingüistas/filólogos
X **Variación diatópica**	Lugar de publicación en vez de lugar de nacimiento	Metadatos precisos (variedades del español, países, lugar de origen del autor) – TEI
X **Variación diastrática**	Investigación manual adicional	Metadatos precisos (detalles biográficos del autor, editores) – TEI
X **Variación diafásica**	CORDE: "tema" (clasificación compleja) CdE: cuatro géneros, dudosa representatividad de cada uno	Metadatos precisos (tipología textual) – TEI

Tabla 3: Estudios de variación sociolingüística con corpus diacrónicos del español: problemas y soluciones. Fuente: Elaboración propia.

Además, los corpus diacrónicos no sirven para la investigación de todos los niveles lingüísticos, o necesitan mucha revisión adicional:

- No son apropiados para estudios grafemáticos y fonético-fonológicos, ya que los textos incluidos poseen criterios de edición muy diversos, lo que hace difícil la especulación de todas las posibles variantes (ej.: *haber, aver, auer, haver*, etc.). La Red Internacional CHARTA (Corpus Hispánico y Americano en la Red: Textos Antiguos) constituye un ejemplo modélico, pues todos sus textos se están editando con los mismos criterios, creados y usados por los participantes del proyecto.
- Con respecto a los estudios morfosintácticos, son destacables los avances que se han realizado a partir del CORDE, que solo permite la búsqueda de secuencias de palabras exactas (por ejemplo: *que su*, para estudiar el quesuismo), debido a la arquitectura del corpus, basada en el servicio de indexación de Microsoft (Davies, 2009: 165). Gracias a la lematización y etiquetación morfosintáctica del CdE y del CDH, es posible realizar búsquedas por categorías gramaticales. No obstante, solo CdE

permite información gramatical detallada. Por ejemplo, en el caso de los verbos, se puede realizar una búsqueda teniendo en cuenta las desinencias de persona, número, tiempo, aspecto y modo. Sin embargo, la lematización de este corpus se ha llevado a cabo siguiendo métodos estadísticos y sin revisión posterior, por lo que contiene errores considerables. Volviendo al estudio de *vos*, al buscar todos los verbos que contengan desinencias de segunda persona del plural (a través de [v*2p]), obtenemos como uno de los resultados el sustantivo *cibdad* (variante medieval de *ciudad*), con 175 ocurrencias.

- Por otra parte, los estudios léxico-semánticos son, junto con los diacrónicos, los más usados por los investigadores. De hecho, como señaló en mi entrevista uno de los coordinadores de los corpus de la RAE, estos están diseñados para servir de base al *Nuevo Diccionario Histórico* y, por tanto, son más apropiados para la investigación léxico-semántica. Sin embargo, si el corpus no está totalmente lematizado y revisado, es necesario un gran esfuerzo para especular las posibles variantes con las que podría estar escrita una palabra.

Nivel lingüístico	Análisis con corpus	Soluciones
X **Grafemático Fonético-Fonológico**	Diversos criterios edición Difícil: especular variantes	Criterios de edición comunes/explícitos Lematización revisada (Freeling – EAGLES)
X **Morfosintáctico**	Limitado a palabras o secuencias de palabras (*que su/cuyo*)	Lematización revisada Anotación morfosintáctica (categoría gramatical, subcategorías) (Freeling – EAGLES)
✔ **Léxico-Semántico**	Pero… esfuerzo de especulación de variantes	Lematización revisada (Freeling – EAGLES)

Tabla 4: Estudios de los distintos niveles lingüísticos con corpus diacrónicos del español: problemas y soluciones. Fuente: Elaboración propia.

En definitiva, la lematización y la anotación morfosintáctica de un corpus (con revisión y corrección posteriores) son esenciales para llevar a cabo estudios lingüísticos en todos los niveles. Para ello, también debe usarse un estándar internacional como EAGLES (Expert Advisory Group on Language Engineering Standards), en el que se basa el etiquetador Freeling, que ha sido adaptado a textos españoles anteriores al siglo XX. El uso de estándares internacionales es fundamental para programas y recursos digitales del siglo XXI: es imprescindible para evitar la dependencia del autor del programa, y sobre todo para la compatibilidad de programas y recursos, así como por motivos de preservación del enorme trabajo y esfuerzo que supone la etiquetación lingüística. Sin embargo, la falta de una interfaz amigable dificulta el uso de Freeling.

Un corpus digno de imitación, que ha sido etiquetado con TEI y lematizado/anotado morfosintácticamente con Freeling (basado en EAGLES) –con posterior revisión filológica– es Post Scriptum. Se trata de un corpus de cartas de español y portugués de la Edad Moderna. Permite realizar estudios fiables de todos los niveles lingüísticos, así como estudios de sociolingüística histórica, con opciones de búsqueda multivariable.

Análisis de lematizadores y etiquetadores morfosintácticos del español anterior al siglo XX

Dada la importancia de los lematizadores y etiquetadores, he evaluado[2] Freeling (en constante actualización) y AyDA (años 80) –ambos gratuitos–, con el permiso de sus autores, aplicándolos a textos de diferentes periodos del español, obteniendo los resultados que se observan en el siguiente gráfico, que son similares a los

[2] Para un análisis más detallado, véase Díaz-Bravo (2015: 383-389).

resultados procedentes de los análisis llevados a cabo por los propios creadores de cada programa (Sánchez, Boleda y Pradó, 2011: 1; Capelli y Saba, 2003: 885):

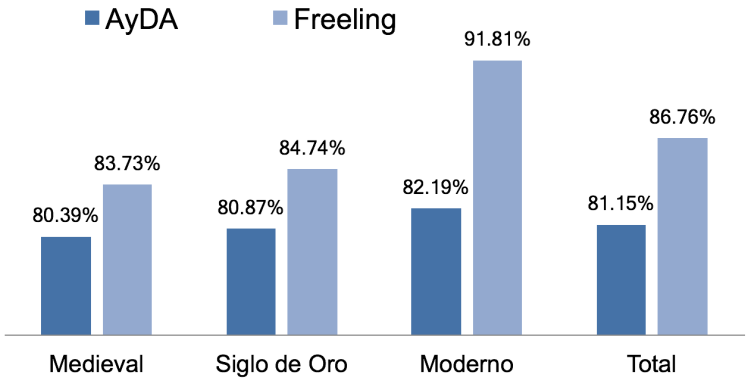


Figura 6: Resultados de la evaluación de lematizadores y etiquetadores morfosintácticos (AyDA y Freeling). Fuente: Elaboración propia

Asimismo, he analizado el lematizador Bconcord (de B. Horcajada). Aunque no obtuve el permiso para adquirirlo y probarlo con textos de diferentes épocas, sí que pude evaluar su interfaz, usabilidad y tecnología (Díaz-Bravo, 2015: 383-389). Su principal inconveniente es que no sigue estándares internacionales.

A pesar de que los corpus lematizados y anotados morfosintácticamente se consideran esenciales para los estudios lingüísticos, existe una escasez de lematizadores del español adaptados al español medieval y clásico, así como de corpus lematizados con revisión posterior. Se pueden argumentar las siguientes razones: la dificultad de desarrollar lematizadores para el español anterior al siglo xx (que posee un alto grado de variación gráfica), así como la falta de investigadores que combinen los conocimientos de Lingüística y Filología con Humanidades Digitales y Lingüística Computacional. Por otra parte, los lematizadores para el español

anterior al siglo xx deberían mejorarse, ya que el número de errores obtenidos de cada uno es todavía significativo (y el resultado sería mucho más negativo si se aplicara a transcripciones paleográficas). El análisis de las necesidades de los investigadores demuestra que la precisión de los datos es más importante para ellos que la cantidad. A pesar de la incompatibilidad entre estos lematizadores, sería posible combinar algunos de los diccionarios que forman parte de ellos, así como incorporar nuevos *scripts* nacidos de la observación de las reglas lógicas en que se basan. Asimismo, deben destacarse las técnicas de aprendizaje automático, gracias a las que el nuevo *input* corregido sirve para mejorar la precisión del lematizador. Para ello, la colaboración y la unión de esfuerzos resultan esenciales.

Conclusiones

En este trabajo he intentado demostrar –a través del caso práctico del estudio de *vos*– que, aunque los grandes corpus diacrónicos del español son recursos de enorme utilidad para los investigadores de historia de la lengua española, deben mejorar significativamente en aspectos textuales y, sobre todo, digitales, para que se pueda aprovechar todo su potencial y para permitir estudios precisos de todos los niveles lingüísticos y de todos los tipos de variación lingüística.

Tras este breve análisis, quisiera destacar las siguientes conclusiones finales:

- La importancia de una planificación detallada de todas las fases implicadas en la creación de un recurso digital y de un corpus lingüístico, teniendo en cuenta, entre otros aspectos, las necesidades de sus usuarios, el *output*, la recuperación de la información y la compatibilidad con otros recursos digitales y programas informáticos.

- La necesidad de corpus lematizados con opciones de búsqueda avanzada**,** anotación lingüística e interfaces amigables; así como de mejorar los lematizadores y etiquetadores morfosintácticos.
- La relevancia de los estándares internacionales (EAGLES, TEI), por razones de trasferibilidad y preservación. Por ello, debe subrayarse la conveniencia de usar Freeling para la lematización y anotación lingüística de corpus del español. No obstante, la revisión y corrección de los datos lematizados y etiquetados con corpus que han sido sometidos a un tratamiento automático es fundamental para que los resultados ofrecidos sean fiables.
- Un mayor entendimiento entre disciplinas (Lingüística Computacional, Humanidades Digitales, Historia de la Lengua Española).
- Para todo ello, son fundamentales la colaboración, la unión de esfuerzos y de criterios.

Referencias bibliográficas

Capelli, G. y Saba, A. (2003). MORFSIN and AyDA: two systems for analyzing modern and old Spanish. En A. Zampolli, N. Calzolari y L. Cignoni (Eds.), *Linguistica Computazionale*, vol. XVIII-XIX (pp. 865-900). Pisa: Istituti Editoriale e Poligrafici Internazionali.

CLUL (Ed.) (2014). *P.S. Post Scriptum. Arquivo Digital de Escritura Quotidiana em Portugal e Espanha na Época Moderna*. Recuperado de http://ps.clul.ul.pt [12/02/2018]

Davies, M. (2009). Creating useful corpora: A comparison of CORDE, the Corpus del Español, and the Corpus do Português. En A. Enrique-Arias (Ed.), *Diacronía de*

las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus (pp. 137-166). Madrid/Frankfurt: Iberoamericana/Vervuert, Lingüística Iberoamericana, vol. 37.

—(2001). *Corpus del Español (CdE)*. Recuperado de http://www.corpusdelespanol.org [12/02/2018]

Díaz-Bravo, R. (en prensa). *Francisco Delicado,* Retrato de la Loçana andaluza*: A Critical Edition with Introduction and Notes*. Cambridge: Modern Humanities Research Association, Volumen 56.

—(2015). Herramientas computacionales aplicadas al estudio de la historia de la lengua española. En J. P. Sánchez Méndez y M. de la Torre Viorica Codita (Eds.), *Temas, problemas y métodos para la edición y el estudio de documentos hispánicos antiguos* (pp. 377-393). Valencia: Tirant Humanidades.

—(2010). *Estudio de la oralidad en el* Retrato de la Loçana andaluza *(Roma, 1524)*, (Tesis Doctoral). Recuperado de http://riuma.uma.es/xmlui/handle/10630/4575.

—(2007). Transición del sistema de tratamientos medieval al del Siglo de Oro. En *Actas del XXIV Congreso Internacional de AESLA (Asociación Española de Lingüística Aplicada), Congreso Nacional de Lingüística Aplicada: Aprendizaje de Lenguas, Uso del Lenguaje y modelación cognitiva* [recurso electrónico]: perspectivas aplicadas entre disciplinas, Madrid, UNED, 30 de marzo-1 de abril 2006.

Koch, P. y Oesterreicher, W. (2007). *Lengua hablada en la Romania: español, francés, italiano* (versión española de

Araceli López Serena; revisada, actualizada y ampliada por los autores). Madrid: Gredos.

RAE. *Corpus del Nuevo Diccionario Histórico del Español (CDH)*. Recuperado de http://web.frl.es/CNDHE/view/inicioExterno.view [12/02/2018]

—*Corpus Diacrónico del Español (CORDE)*. Recuperado de http://corpus.rae.es/cordenet.html [12/02/2018]

—*Corpus del Español del Siglo XXI. (CORPES)*. Recuperado de http://web.frl.es/CORPES/view/inicioExterno.view [12/02/2018]

Sánchez-Marco, C., Boleda, G., Padró, Ll. (junio, 2011). Extending the tool, or how to annotate historical languages varieties (pp. 1-9). En *Proceedings of the 5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* LaTeCH, The Association for Computational Linguistics, Portland, Oregon, USA. Recuperado de http://www.aclweb.org/anthology/W11-1500.

Torruella, J. *Corpus Informatizat del Català Antic (CICA)*. Recuperado de http://www.cica.cat/ [12/02/2018]

Breve bio de la autora

Rocío Díaz Bravo holds a BA in Hispanic Philology (2004) and a PhD in Spanish Linguistics (2009) from the University of Málaga. She has completed Master's degrees in Teaching Spanish as a Foreign Language (UMA, 2007) and in Digital Humanities

(King's College London, 2011). Her PhD thesis is a linguistic study of the orality of the book *Retrato de la Loçana andaluza* (Delicado,1530?). She is currently working as a Lecturer in Spanish Linguistics at the University of Granada. She has taught Spanish Language and Linguistics in four European countries, especially in the UK (e.g., University of Cambridge and University College London, where she worked as the Language Coordinator of Spanish Language). Her areas of interest include the following: History of the Spanish Language, Varieties of Spanish, Teaching and Learning Spanish. She specialized in the use of new technologies applied to both teaching and research of Spanish Language. Her research in Digital Humanities focuses in the analysis of digital resources and computational tools for the study of the History of the Spanish Language.

Rocío Díaz Bravo es Licenciada en Filología Hispánica (2004) y Doctora en Lengua Española (2009) por la Universidad de Málaga. Posee estudios de Máster en la enseñanza de ELE (UMA, 2007) y en Humanidades Digitales (King's College London, 2011). Su investigación realizó un estudio lingüístico de la oralidad en el *Retrato de la Loçana andaluza* (Delicado, 1530?). Actualmente es Profesora de Lengua Española en la Universidad de Granada. Ha sido profesora de Lengua Española en cuatro países europeos, especialmente en Reino Unido (por ejemplo, Universidad de Cambridge y University College London, donde trabajó como Coordinadora de Lengua Española). Sus líneas de investigación principales son las siguientes: Historia de la Lengua Española, Variedades del Español, Enseñanza-aprendizaje de ELE. Se ha especializado en el uso de las nuevas tecnologías aplicadas tanto a la investigación como a la enseñanza de la lengua española.. Su investigación de Humanidades Digitales se centra en el análisis de recursos digitales y herramientas computacionales aplicadas al estudio de la Historia de la Lengua Española.