Revista de Variación y Cambio Lingüístico (3) 2026
ISSN: 3020-9854
Doi: https://doi.org/10.30827/3020.9854rvcl.3.2026.37868

Los corpus y la lingüística del corpus del español: desarrollos recientes*

Los corpus lingüísticos, y la lingüística de corpus (LC) como disciplina y metodología, han supuesto una "revolución instrumental" (Rojo, 2021, p. 44) en la investigación lingüística, al situar los datos empíricos como objeto central de análisis y superar enfoques más tradicionales, frecuentemente basados en modelos intuitivos, que evocan al llamado "armchair linguist" (Fillmore, 1992, p. 35) o "lingüista de sillón". Los avances tecnológicos y computacionales de las últimas décadas han hecho posible, además, la recopilación, almacenamiento y explotación automatizada de grandes volúmenes de datos (*vid*. Parodi, 2008; Rojo, 2015, para más información). Los corpus permiten observar el uso de la lengua en su contexto de producción natural, identificar frecuencias léxicas, patrones gramaticales y discursivos recurrentes, y establecer correlaciones entre distintos fenómenos lingüísticos. El acceso a corpus de referencia, orales y multimodales, con distintos tipos y niveles de anotación, ha permitido el análisis sistemático y cuantificable de datos textuales reales, contribuyendo a la modernización de la investigación lingüística. En palabras de Parodi (2008, p. 97): "[d]esde la LC, con el despuntar del medio siglo XX, son muchos los lingüistas que anhelan indagar el uso lingüístico, tal como es producido, comunicado y comprendido entre hablantes/escribientes y oyentes/lectores reales y en situaciones concretas y particulares". Así pues, hoy en día no se cuestiona que los corpus, junto con la LC, han contribuido decisivamente a la descripción empírica de las lenguas (Parodi, 2010) y han sentado las bases para la elaboración de gramáticas, diccionarios y otros recursos lingüísticos de referencia.

Para comprender qué entendemos por corpus lingüísticos, partimos de la definición que proporciona Sinclair (2005, p. 16), quien los describe como "collection[s] of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research". Briz (2012, p. 121), por su parte, resume de esta manera las características de los "corpus propiamente dichos":

> — Están formados por muestras de habla obtenidas o recogidas en su contexto natural de enunciación. [...]
> — Son los propios investigadores, los grupos de trabajo o personal formado para tal fin quienes recolectan esas muestras.
> — Son de acceso directo al objetivo.
> — Se presentan directamente en formato textual. Por ello hablamos de corpus propiamente dichos, porque permiten siempre el acceso directo a los textos completos [...] necesario cuando uno realiza estudios sociolingüísticos, sociopragmáticos o pragmalingüísticos.
> — Tienen objetivos precisos y, por tanto, límites más definidos, lo cual no impide que puedan servir a otros fines (incluso, no previstos).
> — Son muestras de tipología homogénea (lo que también es consecuencia de lo anterior).

* Quiero expresar mi agradecimiento a *Variación. Revista de variación y cambio lingüístico* por la acogida de este número temático y, de manera particular, a la directora de la revista, la profesora Doina Repede, por su implicación, su constante disposición y el apoyo prestado a lo largo de todo el proceso editorial (desde la revisión de los artículos hasta su maquetación), que han sido fundamentales para la materialización de este proyecto. También les doy las gracias a todas las personas que han contribuido a la realización de este número, tanto mediante la autoría de los trabajos aquí reunidos como mediante la altruista labor de revisión, imprescindible para garantizar su calidad científica.

— Los frutos principales obtenidos de su explotación están en relación con esos objetivos.
— Pueden estar informatizados o no. No siempre están marcados, ni tienen acceso electrónico por concordancias (aunque sería deseable y, sobre todo si son macro-corpus, muy necesario).
— Suele haber muestras impresas.
— Y suelen partir de hipótesis previas [...]

Con esta perspectiva, los corpus se conciben como bases de datos intencionadamente diseñadas y metodológicamente controladas, pues la recogida de las muestras no solo no es arbitraria, sino que responde a unos objetivos de investigación predefinidos. En este sentido, se enfatiza el carácter empírico de los datos que contienen los corpus, procedentes de situaciones comunicativas reales. El acceso a los textos completos por los que están constituidos los corpus, y a su contexto de producción natural, se revela esencial, además, para la interpretación de los fenómenos objeto de análisis en disciplinas como la sociolingüística o la pragmática. Esta descripción, que también destaca el papel activo de los investigadores en la construcción de un corpus, contrasta con algunos enfoques más recientes en los que los corpus tienden a ser más amplios, heterogéneos y menos restringidos por hipótesis iniciales, como los que recopilan masivamente textos de internet[1]. Por otro lado, ya no parece posible contemplar la posibilidad la inclusión sistemática de muestras impresas, dada la generalización de los grandes corpus en formato electrónico. Por tanto, la caracterización que proponía Briz (2012) refleja los rápidos cambios que han experimentado los corpus a lo largo de los últimos años, dejando a un lado su carácter más "artesanal" y dando cabida a una mayor automatización en la recopilación de los datos.

La creciente especialización en los corpus del español es un aspecto también destacable en su evolución a lo largo de los últimos años, que se evidencia en la creación de corpus de temáticas específicas o de ámbitos más restringidos, como corpus interacciones digitales (ej., CODICE), de actos de habla (ej., COREMAH), de géneros especializados (ej., HUMCOR, HUMNET), de variedades geolectales y sociolectales concretas (ej., PRESEEA, ESLORA), o incluso, de hablantes no nativos, los llamados corpus de aprendices o aprendientes (*vid*. Rojo *et al*., 2022). Esta diversidad, a la que ya apuntaban hace más de 15 años Briz y Albelda (2009) y que se ilustra en los corpus que se presentan en este número temático, se explica por la necesidad de conjugar las nuevas formas de comunicación del presente con el interés por el estudio de ciertas cuestiones lingüísticas difícilmente observables en corpus tradicionales, como la variación gráfica o los nuevos géneros discursivos (*vid*. como muestra, Mancera Rueda y Pano Alamán, 2013; Sampietro, 2023; Vela Delfa y Cantamutto, 2024). Así pues, el auge de la comunicación en entornos digitales ha impulsado la creación de corpus de redes sociales, foros, mensajería instantánea y otros medios virtuales, abriendo paso al análisis de este tipo de fenómenos.

A pesar de los avances tecnológicos en la compilación de corpus, la LC del español sigue enfrentándose a importantes desafíos. Uno de los principales tiene que ver con la representatividad de las muestras, ya que, más allá de algunos corpus panhispánicos, existen desequilibrios en la presencia de geolectos, registros, contextos comunicativos y grupos sociales. Así, por ejemplo, Madrid cuenta con un ingente volumen de datos de estudio (ej., PRESEEA Madrid, CORMA, CORLEC, COLA, Corpus de Habla Juvenil, MADSex, entre otros), mucho mayor con respecto a otras comunidades cercanas como las dos Castillas o Aragón, lo

[1] Muestra de ello es la plataforma Sketch Engine, que alberga numerosos corpus en línea y recursos colaborativos, como TenTen Corpora (https://www.sketchengine.eu/documentation/tenten-corpora/) por citar alguno.

que limita la posibilidad de obtener una visión más abarcadora de la diversidad lingüística del centro peninsular o de establecer comparaciones. A ello se suma la todavía limitada disponibilidad de corpus específicamente diseñados para el análisis de fenómenos pragmáticos y discursivos, ya que siguen siendo escasos los recursos que permiten estudiar de manera sistemática cuestiones como las estrategias de cortesía, la gestión de la interacción, la construcción de identidades discursivas, el análisis prosódico o la multimodalidad.

Así, en la actualidad, aunque son numerosas las disciplinas que se benefician de los corpus, como la lingüística computacional, la lexicografía o la morfosintaxis, el aprovechamiento de estos recursos continúa siendo desigual: mientras que el análisis de fenómenos morfosintácticos y léxicos ocupa una posición privilegiada en la investigación con corpus, otras áreas, como la pragmática, la fonética o el análisis del discurso, han recibido una atención comparativamente menor (Sampedro Mella, 2021a). Este desequilibrio responde, en cierta medida, al papel histórico de la tradición gramatical en los estudios lingüísticos, pero también a los sistemas de etiquetado morfosintáctico presentes en la mayoría de los corpus anotados. Dichos sistemas facilitan la recuperación de palabras y sintagmas, categorías gramaticales o, en muchos casos, lemas, a través de búsquedas por concordancias; pero no de contenidos pragmáticos, discursivos o textuales; fenómenos fonéticos o prosódicos, o información multimodal. Por consiguiente, para llevar a cabo estudios sobre este tipo de cuestiones con datos de corpus, suele ser preciso realizar búsquedas específicas por formas lingüísticas concretas (O'Keffe, 2018), o bien revisar manualmente los materiales por los que están constituidos los corpus (Sampedro Mella, 2021a).

Aun cuando los sistemas de anotación lingüística y las herramientas de procesamiento automático han experimentado grandes avances computacionales en las últimas décadas, las mejoras han repercutido, sobre todo, en la anotación morfosintáctica, principal nivel de enriquecimiento de los corpus durante décadas. Ahora bien, como explica García-Miguel (2022, pp. 18-19), cualquier nivel de análisis lingüístico es susceptible de anotación: el fonológico (transcripción fonética, límites silábicos, rasgos prosódicos, etc.), el ortográfico, el morfológico, el léxico-gramatical (lematización, clases de palabras, rasgos morfosintácticos), el sintáctico (constituyentes, relaciones de dependencia, funciones…), el semántico (campos semánticos, desambiguación léxica…), el textual-discursivo (relaciones anafóricas, tema/rema, estructura del discurso…) o el pragmático (actos de habla, análisis de sentimientos y opiniones, roles discursivos, etc.). Si bien se observa una progresiva incorporación de niveles de análisis semántico, pragmático y discursivo en algunos corpus del español (ej., COREMAH, OBSERVAHUMOR), que permiten la recuperación de estos datos, este desarrollo plantea a su vez nuevos retos metodológicos: la calidad de la anotación, la necesidad de establecer unos criterios homogéneos de aplicación y la validación manual de los resultados automáticos. El desarrollo de técnicas automáticas de lematización, etiquetado y análisis sintáctico incrementa considerablemente la capacidad de explotación de grandes cantidades de información. No obstante, la anotación automática no siempre es posible (todavía), en especial en aquellos niveles sujetos a la interpretación, como la pragmática (Weisser, 2015; Milà-Garcia, 2018). La anotación manual, por su parte, tampoco es factible en corpus con ingentes cantidades de datos (cf. García-Miguel, 2022), lo que de nuevo limita las posibilidades de incorporar otros niveles de anotación y de recuperar automáticamente fenómenos alejados de las concordancias que permiten las búsquedas.

Por último, merece una mención especial el reciente, y creciente, empleo de los corpus en la enseñanza del español como lengua extranjera (ELE), impulsado por la compilación de corpus de aprendices, como el Corpus de Aprendices de Español (CAES), el Corpus Escrito del

Español L2 (CEDEL2), el Corpus Español Multimodal de Actos de Habla (COREMAH) o el Davis Corpus of Written Spanish, L2 and Heritage Speakers (COWS-L2H), entre otros (*vid.* Rojo *et al.*, 2022, para una nómina y descripción de estas herramientas), así como por el diseño de propuestas didácticas y estudios empíricos específicos[2]. En paralelo a los estudios de la variedad nativa descritos arriba, estas nuevas investigaciones sobre la adquisición del español, junto con las propuestas didácticas, se han orientado principalmente al análisis y la enseñanza de cuestiones gramaticales y léxicas. Sin embargo, en el ámbito del enfoque comunicativo por tareas, los corpus pueden ser utilizados con gran provecho para el estudio de otras cuestiones lingüísticas, pragmáticas, discursivas y fónicas, que favorezcan el contacto del aprendiz con el uso real de la lengua (Ballestero de Celis y Sampedro Mella, 2021; Sampedro Mella y Rodríguez Domínguez, 2026).

Con el objetivo de contribuir al desarrollo de la LC del español y de fomentar la utilización de corpus en disciplinas en las que su aprovechamiento sigue siendo más limitado por las razones antes apuntadas, surge este número temático, que recopila quince artículos sobre temáticas diversas, con enfoques y objetivos distintos, pero unidos por un denominador común: los corpus del español. Se presentan tanto investigaciones inéditas como nuevos corpus y funcionalidades, desarrollados por distintos equipos, o individualmente, en diferentes universidades europeas. Dada la heterogeneidad de las contribuciones incluidas, se ha optado por organizarlas en torno a dos bloques.

El primer bloque, titulado **Avances en los corpus del español**, tiene como fin mostrar algunas de las innovaciones que han experimentado varios corpus consolidados del español (plataforma OBSERVAHUMOR, ESLORA, CEDEL2), así como dar a conocer nuevos proyectos y corpus inéditos desarrollados en el ámbito europeo (DESEO, COFRE, CORAV, COMMAR). Las contribuciones aquí reunidas ponen de relieve la sofisticación progresiva de los sistemas de anotación y de las herramientas de recogida y procesamiento de los datos. Asimismo, destaca la creación de nuevos corpus a partir de interacciones digitales, o mediante la incorporación de muestras orales de hablantes nativos y de herencia, a través de conversaciones y entrevistas. En conjunto, estos corpus reflejan un nuevo rumbo en la LC, orientado a las formas de comunicación digital más recientes, a otros géneros textuales, como el humor; y a la creciente diversidad de hablantes de español, en un afán por abarcar los múltiples y variados usos lingüísticos actuales.

Inicia la primera parte de este bloque, Leonor Ruiz Gurillo, con el artículo "Digitalizar, compilar y etiquetar corpus de humor. La experiencia con la plataforma OBSERVAHUMOR", en el que, por una parte, describe esta plataforma y los seis corpus que alberga actualmente. Por otra, tomando como referencia el corpus CHILDHUM, ilustra el proceso íntegro de creación de este recurso, que contiene narraciones humorísticas de niños y niñas entre 8 y 12 años, desde la recogida de los datos hasta su publicación en la web, así como la anotación pragmática de distintos fenómenos discursivos y su posterior explotación. También en el ámbito de la anotación pragmática, el segundo artículo, "Dando voz a la pragmática: creación de un sistema de anotación de actos de habla en el corpus ESLORA" (María Sampedro Mella), describe el proceso de anotación de una selección de actos de habla en ESLORA, destacando la necesidad de conciliar la variabilidad del uso lingüístico con la estructuración del sistema computacional, así como los retos encontrados durante este proceso y las soluciones

[2] Sirvan de muestras los siguientes: Mas Álvarez y Gil González (2018), Abad Castelló (2019), Abad Castelló y Álvarez Braz (2019), Marcos Miguel (2020), Sampedro Mella (2021b, 2022, 2024), Álvarez Ejzenberg (2021), García Márkina y Valdez (2021), Repede (2025), Sampedro Mella y Rodríguez Domínguez (2026), Santana y Repede (en prensa), Cabedo Nebot, Fernández de Molina y Repede (2026).

adoptadas. Cierra esta primera parte "Aportaciones de la versión 3 del corpus CEDEL2 de español L2/LE" (Cristóbal Lozano, Nobuo Ignacio López-Sako y Jorge Montaño), que traza la evolución de este corpus de aprendices desde sus inicios en 2006 hasta su versión actual, que incorpora muestras de interlengua orales y escritas de aprendices con quince lenguas maternas distintas, y varios corpus nativos. Los autores exponen los cambios que ha experimentado tanto la recogida de los materiales como su diseño computacional, que permite realizar búsquedas de diferentes tipos.

La segunda de este bloque parte contiene cuatro artículos en los que se presentan sendos corpus novedosos e inéditos, con notorias diferencias entre sí, que reflejan los cambios sociales y digitales actuales, y cómo los nuevos corpus se han adaptado a estos escenarios. Vicente J. Marcet Rodríguez, en "Un corpus llamado DESEO. Análisis lingüístico de los perfiles de la aplicación de contactos homoeróticos Grindr", nos acerca a este corpus, a partir de un riguroso análisis lingüístico de sus muestras, que, aun siendo escritas, se sitúan en una frontera entre la oralidad y la escritura. Prosigue el artículo "COFRE – *Corpus of Online Forum Replies*. A Trilingual TripAdvisor Corpus for the Study of Address Forms and Beyond" (María Sampedro Mella, Louis-Amélie Cougnon y Barbara De Cock), sobre este corpus trilingüe (francés, español y portugués) basado en interacciones de TripAdvisor, recopiladas de foros de esta plataforma de Bélgica, Francia, España y Portugal. A continuación, en "Corpus de la lengua hablada de Ávila (CORAV): presentación y análisis de la construcción discursiva de problemas sociales", Vanesa Álvarez Rosa describe el corpus CORAV (PRESEEA Ávila), a partir de los criterios metodológicos que han guiado su diseño y su construcción, y muestra un ejemplo de explotación de los materiales que contiene, mediante un análisis del discurso de tres entrevistadas. Finalmente, el artículo "Contacto árabe-español: diseño y compilación del Corpus Gradia de Mujeres Marroquíes de Cataluña (COMMAR)" (Samia Aderdouch Derdouch) presenta este corpus, que reúne producciones orales de 100 mujeres con el árabe marroquí como lengua de herencia o de inmigración, escolarizadas en España o Marruecos, junto con algunos resultados preliminares del análisis de varios fenómenos morfosintácticos y de contacto léxico.

El segundo bloque, denominado **Explotación de corpus**, tiene como fin mostrar algunas investigaciones empíricas novedosas realizadas a partir de corpus existentes o de nueva creación. También se incluyen acercamientos interdisciplinarios entre la LC y otras disciplinas, como el paisaje lingüístico o la disponibilidad léxica, sin olvidar las aplicaciones didácticas de los corpus en la enseñanza de ELE. Los estudios aquí recogidos detallan el proceso de explotación de los datos lingüísticos de los corpus mediante análisis cuantitativos y cualitativos, que permiten identificar pautas de uso, frecuencias y variación contextual. De este modo, estos trabajos aportan nuevos resultados, y a su vez, por su enfoque, pueden ser también replicables metodológicamente.

Por un lado, se presentan cuatro investigaciones elaboradas a partir de corpus existentes, con una sólida tradición, o de nuevos corpus. Primero, Alba Fernández Sanmartín, en "La entrevista sociolingüística en el corpus ESLORA: análisis pragmático-discursivo de la interacción entrevistador–entrevistado", examina, desde una doble perspectiva cuantitativa y cualitativa, las intervenciones de los entrevistadores en una selección de entrevistas de corte tradicional frente a otras libres, más cercanas a la conversación, y las variaciones que producen en la interacción. Prosigue el artículo "Agent-Defocusing Mechanisms in French and Spanish: A Contrastive Corpus-Based Study" (Emeline Pierre), que examina las estrategias discursivas de desfocalización del agente en francés y español en una selección de corpus comparables con diferentes niveles de formalidad: ESLORA, CLAPI, CIEL-F y EUROPARL. En

"Construcción y explotación de un corpus de prensa catalana en castellano (1800–1849): aplicación a las perífrasis modales de infinitivo" (Yujian Han), se detalla el proceso de compilación, digitalización y transcripción automatizada del corpus Gradia_prensa_SIGLO XIX, que sirve a este autor como base para el estudio de las perífrasis modales en este periodo. El artículo "Del ALPI a los corpus actuales: el *heheo* en la evolución fonética del sur castellanoleonés", de Gonzalo Francisco Sánchez, describe este cambio fonético en varios corpus, desde el ALPI hasta datos actuales propios, y subraya la importancia del trabajo de campo en zonas rurales para el estudio de fenómenos fonéticos en variación.

Por último, cierran este bloque cuatro trabajos que reflejan la estrecha relación de la LC con otras disciplinas, incluso con un desarrollo más reciente en el ámbito lingüístico. El artículo "Corpus y paisajes lingüísticos. Una revisión crítica", de Carmen Fernández Juncal, examina el papel de los corpus lingüísticos en los estudios sobre paisaje lingüístico, así como la elaboración de corpus específicos de paisajes lingüísticos y las dificultades metodológicas que plantea, debidas entre otras causas, al carácter efímero de las muestras; también se analizan las posibilidades de estos corpus en la investigación y en la enseñanza. A continuación, el artículo "De la disponibilidad léxica al aula de ELE: Corpus, análisis de errores y orientaciones didácticas" (Roberto Rubio Sánchez) muestra cómo los estudios de disponibilidad léxica pueden contribuir a la detección temprana de errores en aprendices de ELE de una determinada lengua materna; además, el autor ilustra, por medio de varios ejemplos, cómo esta información puede utilizarse con fines pedagógicos, para la elaboración de materiales didácticos. Prosigue el estudio "Bases para la construcción de un corpus a partir de un catálogo de sustantivos (semi)transparentes con género gramatical diferente en español y en francés", de Fabiana Álvarez-Ejzenberg y Graciela Villanueva, que recoge un elenco de sustantivos cognados con género diferente en ambas lenguas (ej., *banco/banque*, *valor/valeur*), recopilado manualmente por ambas autoras, así como sus posibilidades de desarrollo en forma de corpus y en la didáctica del español y del francés como lenguas extranjeras. Finalmente, cierra este segundo bloque, y el volumen, el artículo "Corpus para la enseñanza de la variación lingüística en ELE: tres ejemplos para el aula" (Nausica Marcos Miguel). En él se ilustra el potencial pedagógico de los corpus en la enseñanza de la variación lingüística, esencialmente dialectal y léxica, mediante el diseño de actividades basadas en distintos corpus, entre ellos el CORPES XXI (RAE, en línea) y el Corpus del Español (Davies, 2002, 2006).

Los trabajos aquí reunidos reflejan, por un lado, las innovaciones y mejoras tecnológicas que han experimentado algunos corpus del español, así como la aparición de nuevos corpus con muestras y formatos variados, que evidencian la constante evolución de estos recursos y su capacidad para adaptarse a nuevos escenarios comunicativos y tecnológicos. Por otro lado, los estudios empíricos que se presentan ofrecen una visión actualizada de la LC como disciplina y metodología que, en diálogo constante con otras disciplinas lingüísticas (fonética, morfosintaxis, pragmática, análisis del discurso, entre muchas otras), se consolida como un soporte fundamental para la investigación empírica. En conjunto, mediante enfoques diversos, estos estudios son un reflejo de la versatilidad y el potencial de los corpus como herramientas de observación, análisis y descripción de fenómenos lingüísticos y discursivos.

En cuanto a las perspectivas futuras, el desarrollo de los corpus parece estar estrechamente vinculado a los avances en el campo de la inteligencia artificial (IA), que están transformando la manera en que se compila, procesa y analiza la información, por un lado; y se investigan y se estudian las lenguas, por otro. Cabe, pues, esperar una mayor integración entre la LC, el procesamiento del lenguaje natural y la IA, así como una expansión de corpus más dinámicos y actualizables en tiempo real (Incelli, 2025). La automatización en la recopilación de datos y

en el procesamiento de la información sin duda abre nuevas vías de investigación, al tiempo que plantea nuevos retos metodológicos y éticos. En este contexto, la LC, tal como se concibe actualmente, se enfrenta a una renovación profunda, esperemos que sin renunciar al rigor que ha caracterizado siempre su evolución y su desarrollo.

María Sampedro Mella (editora invitada)
Université catholique de Louvain

REFERENCIAS

Abad Castelló, M. (2019). Uso de corpus lingüísticos por y para profesores de español como lengua extranjera. *Revista Electrónica de Didáctica ELE*, *31*, 148-167.

Abad Castelló, M. y Álvarez Baz, A. (2022). Aprendizaje basado en datos en español como lengua extranjera: ampliando el ámbito. *Revista Internacional de Lenguas Extranjeras, 16*, 1-20.

Álvarez-Ejzenberg, F. (2021). Análisis contrastivo de las partículas *pues/pois/puis*: empleos discursivos de la forma *pues* en el español como L2. En C. Ballestero de Celis y M. Sampedro Mella (Eds.), *Aportes del CAES a la enseñanza del español como lengua extranjera* (pp. 235-256). Anejo de *Verba, 81*.

Ballestero de Celis, C. y Sampedro Mella, M. (2021). La didáctica del español como lengua extranjera: un acercamiento desde la lingüística de corpus. En C. Ballestero de Celis y M. Sampedro Mella (Eds.), *Aportes del CAES a la enseñanza del español como lengua extranjera* (pp. 9-20). Anejo de *Verba, 81*.

Briz, A. (2012). Los déficits de los corpus orales del español (y de algunos análisis). En T. Jiménez Juliá, B. López Meirama, V. Vázquez Rozas y A. Veiga Rodríguez (Coords.), *Cum corde et in nova grammatica*. *Estudios ofrecidos a Guillermo Rojo* (pp. 115-137). Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico.

Briz, A. y Albelda, M. (2009). Estado actual de los corpus de lengua española hablada y escrita: I+D. En *El español en el mundo. Anuario del Instituto Cervantes 2009*. Instituto Cervantes, 165-226.

Cabedo Nebot, A., Fernández de Molina Ortés, E. y Repede, D. (2026) (Eds.). *Corpus sociolingüísticos y enseñanza de ELE: enfoques y aplicaciones didácticas*. Número especial. *Porta Linguarum*.

Fillmore, C. J. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics". En J. Svartvik (Ed.), *Directions in Corpus Linguistics* (pp. 35-60). Mouton de Gruyter.

García Márkina, Y. y Valdez, C. (2021). El uso de *ser* y *estar* en producciones escritas en L2 con atributo nominal y atributo locativo. En C. Ballestero de Celis y M. Sampedro Mella (Eds.), *Aportes del CAES a la enseñanza del español como lengua extranjera* (pp. 117-146). Anejo de *Verba, 81*.

García-Miguel, J. M. (2022). Lingüística de corpus: de los datos textuales a la teoría lingüística. *Estudios de Lingüística del Español*, *45*, 11-42.

Incelli, E. (2025). Exploring the Future of Corpus Linguistics: Innovations in AI and Social Impact. *International Journal of Mass Communication, 3,* 1-10.

Mancera Rueda, A. y Pano Alamán, A. (2013). *El español coloquial en las redes sociales*. Arco/Libros.

Marcos Miguel, N. (2020). Aplicaciones para la didáctica de ELE del corpus del español. *E-SEDLL, 3*, 32-44.

Mas Álvarez, I. y Gil Martínez, A. (2018). Los corpus de aprendices: un terreno en expansión para la didáctica del español. En M. Ellison (Ed.), *As línguas estrangeiras no ensino superior* (pp. 35-55). Universidade do Porto.

Milà-Garcia A. (2018). Pragmatic annotation for a multi-layered analysis of speech acts: a methodological proposal. *Corpus Pragmatics, 2*(3), 265-287.

O'Keeffe, A. (2018). Corpus-based function-to-form approaches. En A. H. Jucker, K. P. Schneider y W. Bublitz (Eds.), *Methods in Pragmatics (Handbooks of Pragmatics* 10) (pp. 587-618). De Gruyter.

Parodi, G. (2008). Lingüística de corpus. Una introducción al ámbito. *RLA, Revista de Lingüística Teórica y Aplicada, 46*(1), 93-119.
Parodi, G. (2010). *Lingüística de corpus: de la teoría a la empiria*. Iberoamericana.
Repede, D. (2025). La variación geolectal de la atenuación lingüística en ELE. Propuesta didáctica a partir del corpus oral PRESEEA. *Didáctica. Lengua y Literatura*, *37*, 81-100.
Rojo, G. (2015). Sobre los antecedentes de la lingüística de corpus. En A. Álvarez Menéndez *et al*. (Coords.), *Studium grammaticae. Homenaje al Profesor José Antonio Martínez* (pp. 675-689). Universidad de Oviedo.
Rojo, G. (2016). Los corpus textuales del español. En J. Gutiérrez-Rexach (Ed.), *Enciclopedia lingüística hispánica* (pp. 285-296). Routledge.
Rojo, G. (2021). *Introducción a la lingüística de corpus en español*. Routledge.
Rojo, G., I. Palacios, M. Sampedro Mella y A. Marsily (2022). Los corpus de aprendices de ELE: panorama actual y perspectivas futuras. *Journal of Spanish Language Teaching, 9*(2), 174-189.
Sampedro Mella, M. (2021a). DCT vs. corpus orales: reflexiones metodológicas sobre el estudio de los actos de habla. *Pragmalingüística, 29*, 337-395.
Sampedro Mella, M. (2021b). *Estimado Sr.* vs. *Hola hotel*. El análisis contrastivo de la interlengua para la enseñanza de la variación pragmático-discursiva. En V. Marcet Rodríguez, V. Álvarez Rosa y M. Nevot (Eds.), *La variación en español y su enseñanza. Reflexiones y propuestas didácticas* (pp. 134-150). Universidad de Salamanca.
Sampedro Mella, M. (2022). Corpus de aprendices y cuestiones de discurso en la interlengua: un nuevo acercamiento aplicado a la enseñanza. *Foro de profesores de E/LE, 18*, 163-182.
Sampedro Mella, M. (2024). An approach to the lexical ambiguity caused by false cognates in Spanish L2. A corpus-based exploratory study. *Studia Linguistica, 78*(1), 186-205.
Sampietro, A. (2023). El auge de los 'stickers' en WhatsApp y la evolución de la comunicación digital. *Círculo de Lingüística Aplicada a la Comunicación, 94*, 271-285.
Santana, J. y Repede, D. (en prensa). Speaking Sevillian Spanish in the SFL classroom. The PRESEEA-Seville corpus as a teaching resource. En Cabedo Nebot *et al*. (Eds.), Corpus sociolingüísticos y enseñanza de ELE: enfoques y aplicaciones didácticas, *Porta Linguarum*, Número especial.
Sinclair, J. (2005). Corpus and Text- Basic Principles. En M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1-16). Oxbow- Books.
Vela Delfa, C. y Cantamutto, L. (2024). Del emoticono al sticker: tendencias en el uso de graficones en interacciones de WhatsApp en lengua española. *Lengua y Sociedad. Revista de Lingüística Teórica y Aplicada*, *23*(2), 777-790.
Weisser, M. (2015). Speech act annotation. En K. Aijmer y C. Rühlemann (Eds.), *Corpus Pragmatics: A Handbook* (pp. 84-114). Cambridge University Press.

CORPUS CITADOS

ALPI – Heap, D. Atlas Lingüístico de la Península Ibérica (ALPI). http://www.alpi.ca
CAES – Rojo, G. y Palacios, I. (Coords.). Corpus de Aprendices de Español (CAES). https://galvan.usc.es/caes
CEDEL2 – Lozano, C. (Coord.). Corpus Escrito del Español L2. https://cedel2.learnercorpora.com/
CHILDHUM – Corpus CHILDHUM. Plataforma OBSERVAHUMOR. https://www.observahumor.com/es/info-childhum
CIEL-F – Corpus International Ecologique de la Langue Française. http://www.ciel-f.org/
CLAPI – Corpus de la LAngue Parlée en Interaction. http://clapi.ish-lyon.cnrs.fr/V3_Accueil.php?interface_langue=EN

CODICE – Cantamutto, L., Vela Delfa, C. y Boisselier, L. (Coords.). Comunicación Digital: Corpus del Español.
COFRE – Sampedro Mella, M. (Coord.). Corpus of Online Forum Replies. UCLouvain.
COLA – A. Myre Jørgensen (Coord.). Corpus Oral de Lenguaje Adolescente [no disponible]
CORAV – V. Álvarez Rosa. Corpus de la lengua hablada de Ávila.
COREMAH – Vacas Matos, M. (2017). Corpus Español Multimodal de Actos de Habla. https://coremah-1a37f.firebaseapp.com/
CORLEC – Marcos Marín (Coord.). Corpus Oral de Referencia de la Lengua Española Contemporánea. Universidad Autónoma de Madrid. https://www.lllf.uam.es/ESP/Info%20Corlec.html
CORMA – Enghels, R., De Latte, F., Roels, N., Van Den Driessche, N. (2025). El Corpus Oral de Madrid (CORMA). Universidad de Gante. https://www.corma.ugent.be/es/
CORPES XXI – Real Academia Española (en línea). Corpus del Español del Siglo XXI (CORPES XXI) [https://www.rae.es/corpes/]. Versión 1.4.
Corpus de Habla Juvenil – Horcajada, B. y Rodrigo Mateos, J. (Coords.). Corpus de Habla Juvenil. https://www.ucm.es/coralingo/corpus-de-habla-juvenil
Corpus del Español – Davies, M. (2002-). Corpus del Español: Género/Histórico. http://www.corpusdelespanol.org/hist-gen/
Corpus del Español – Davies, M. (2016-). Corpus del Español: Web/Dialectos. http://www.corpusdelespanol.org/web-dial/
COWS-L2H – Davis Corpus of Written Spanish, L2 and Heritage Speakers. https://github.com/ucdaviscl/cowsl2h
DESEO – V. Marcet Rodríguez (Coord.). Descripciones de la Sextimidad Online.
ESLORA – Corpus para el estudio del español oral, v. 2.5. https://eslora.usc.es/
EUROPARL – European Parliament Proceedings Parallel Corpus 1996-2011 https://www.statmt.org/europarl/
GRADIA – Plataforma GRADIA, Gramática y Diacronía. https://gradiadiacronia.wixsite.com/gradia/corpus-gradia
HUMCOR – Repede, D. (en línea). Humcor. Corpus de humor oral multimodal en español. https://humcor.snlt.es/
HUMNET – Repede, D. (en línea). Humnet. Corpus de humor digital en español. https://humnet.snlt.es/
MADSex – Pizarro, A. Corpus Madrileño Oral de la Sexualidad. https://corpora.uclouvain.be/catalog/en/corpus/madsex
OBSERVAHUMOR – Ruiz Gurillo, L. (Dir.ª). (en línea). OBSERVAHUMOR [Plataforma web]. Universidad de Alicante. https://griale.dfelg.ua.es/observahumor/
PRESEEA Madrid – Cestero Mancera, A. M., Molina Martos, I. y Paredes García, F. (Coords.). *La lengua hablada en Madrid*. Universidad de Alcalá.