Nueva Revista de Filología Hispánica (*NRFH*), LXX, 2022, núm. 1, 323-331
ISSN 0185-0121; e-ISSN 2448-6558; DOI: 10.24201/nrfh.v70i1.3792

Marta Blanco, Hella Olbertz y Victoria Vázquez Rozas (eds.), *Corpus y construcciones. Perspectivas hispánicas*. Universidade de Santiago de Compostela, Santiago de Compostela, 2019; 320 pp. (Anexos de *Verba*, 79).

Rebeca Pasillas Mendoza
Universidad Nacional Autónoma de México
rebecapasillas@filos.unam.mx
orcid: 0000-0003-4573-7818

La lingüística de corpus es una disciplina que ofrece herramientas contundentes y datos robustos para el estudio sistemático y empírico de las lenguas naturales que posean registros orales o escritos. En noviembre de 2018, la Facultade de Filoloxía de la Universidade de Santiago de Compostela acogió el seminario "Corpus y construcciones. Perspectivas hispánicas", en que se presentaron trabajos vinculados a los procesos de elaboración de corpus, y al análisis basado en corpus tanto de estructuras gramaticales como de relaciones entre léxico y gramática. Como producto de aquel encuentro, Blanco, Olbertz y Vázquez Rozas editaron este volumen que presenta diez importantes trabajos. Cinco estudios gramaticales con base en datos de corpus forman la primera parte del libro, mientras que la segunda está conformada por otros cinco textos que presentan el diseño y desarrollo de corpus de lengua española y gallega.

El capítulo de Torres-Cacoullos y Travis, "Gramáticas en contacto en un corpus bilingüe" (pp. 13-40), abre la primera parte de esta obra. Las autoras demuestran el gran acierto metodológico que supone emplear un corpus bilingüe en el análisis de la convergencia gramatical y, con ello, cuestionan la hipótesis de que el contacto entre las gramáticas de dos lenguas produzca y explique los cambios que ocurren en una comunidad de habla bilingüe. Para resolver el problema de la falta de evidencia cualitativa y cuantitativa que sustente ese supuesto, las autoras prueban que es necesario, después de identificar un proceso de cambio en la comunidad, describir su condicionamiento lingüístico. Para ello, proponen analizar muestras extraídas de un corpus que *a*) circunscriba puntualmente la comunidad de habla; *b*) muestree sistemáticamente a los miembros de esa comunidad; *c*) obtenga y represente, mediante entrevistas sociolingüísticas, secuencias suficientes de habla espontánea de las dos lenguas

Recepción: 7 de abril de 2021; aceptación: 3 de mayo de 2021.

D.R. © 2022. Nueva Revista de Filología Hispánica
Licencia *Creative Commons Attribution-NonCommercial* (CC BY-NC) 4.0 International

analizadas, y *d*) siga convenciones de transcripción metódica que permitan análisis estandarizados. Para ejemplificar, se presenta el caso del *Corpus bilingüe español-inglés de Nuevo México* (NMSEB), del que se extraen datos de tres estructuras morfosintácticas recurrentes en los estudios de variación interna y contacto español-inglés (perífrasis progresiva, uso de subjuntivo en subordinación y expresión variable del sujeto pronominal). Tras un análisis contrastivo de muestras del NMSEB y de otros corpus monolingües usados como punto de referencia, se prueba muy claramente la continuidad lingüística en estas tres estructuras, es decir, la conservación independiente de las dos gramáticas dentro de la comunidad bilingüe, y no la convergencia gramatical, que se esperaría en un escenario de contacto. En suma, este capítulo muestra indudablemente el uso del NMSEB como caso en que un corpus bilingüe resulta el recurso apropiado para el estudio interno y riguroso de los fenómenos ocurridos en el contacto entre gramáticas.

Granvik presenta el segundo capítulo, titulado "Sobre los orígenes de la construcción encapsuladora en español" (pp. 41-79). El autor entrega un estudio diacrónico del comportamiento de nueve sustantivos abstractos (v.gr. *esperanza*, *idea* o *señal*), que tienen la propiedad funcional de sintetizar, o *encapsular*, la información discursiva contenida en una oración completiva con que se combinan y con que pueden establecer algún grado de identidad experiencial, como en *la idea de empezar la práctica* (p. 43). La investigación se basa en un sesudo análisis de cerca de 7 400 casos extraídos del *Corpus del Nuevo diccionario histórico del español* (CDH) que abarcan el período comprendido entre los siglos XIII y XX. La asignación de valores según tres criterios operativos (determinación de la FN, núcleo del que depende y función sintáctica del N) permite al autor organizar las instancias de esta construcción en una escala de tipicidad, de encapsulador típico a encapsulador marginal.

La gran profundidad histórica del corpus revela cuatro períodos en la diacronía tanto de los nominales encapsuladores como de los esquemas en los que se construyen: N (*de*) *que* + oración; N *de* + infinitivo; N *ser* + oración. El texto señala, describe, deslinda e interpreta pormenorizadamente los dinamismos registrados, por lo cual puede afirmarse que éstos sugieren "que la construcción encapsuladora se va expandiendo cada vez más en la lengua" (p. 67), aunque no se trate de un único fenómeno homogéneo que ejerza siempre la misma función discursivo-textual. No obstante su gran diversidad formal y funcional, Granvik logra reconocer cabalmente el origen medieval de esta construcción y sus pautas de desarrollo —como el incremento

y la especialización de la variedad léxica de encapsuladores, por un lado, y el aumento de los usos típicos y de algunos esquemas, por el otro—, todo lo cual descansa en una muy precisa observación y explotación de la información contenida en el CDH.

El tercer capítulo "*Entre miradas de asombro*: aportaciones de la lingüística de corpus al estudio de una construcción con la preposición *entre*" (pp. 81-119), de López Meirama y Mellado Blanco, presenta un estudio del patrón [*entre* + sustantivo$_{\text{plural/corporal}}$], una de las construcciones preposicionales no fijadas y de alta productividad que ha sido relegada en los estudios tradicionales. A partir del montaje, filtrado manual y análisis de un corpus de 1 169 datos extraídos del CORPES XXI —como *entre risas* o *entre besos*, en combinatoria con verbo—, las autoras ofrecen una descripción holística del prototipo de esta construcción, desde el marco de la gramática de construcciones. Es precisamente ese minucioso análisis cualitativo y cuantitativo basado en corpus el que evidencia los rasgos prototípicos de la construcción, a saber, *a*) su valor temporal-modal y función como predicación secundaria en aposición; *b*) que el nombre plural, escueto y deverbal de su *slot* S pertenece al ámbito de la comunicación o expresión corporal (*aplausos*, *bromas*, *susurros*…); *c*) que coaparece con verbos de comunicación o de desplazamiento (*hablaba entre risas*; *se fue entre lágrimas*); *d*) que tiene valor expresivo y de intensificador; *e*) y que surge en registros elevados y en textos literarios. Asimismo, este estudio detallado de corpus permite a las autoras afinar la descripción formal y funcional de la construcción para reforzar y matizar muy pormenorizadamente el prototipo, cuyas realizaciones periféricas también se presentan, además del papel que ejerce en él la *coerción exocéntrica*, el *continuum* léxico-gramatical en que se ubica y, muy destacadamente, la fijación cognitiva (*entrenchment*) de la construcción. En definitiva, este capítulo es un muy valioso modelo para el estudio sistemático de unidades fraseológicas, por su explicitud metodológica y su replicabilidad.

En el cuarto capítulo, "En torno al concepto de *perfil combinatorio*" (pp. 121-146), Mas Álvarez ofrece un nutrido panorama de diversos estudios —Hanks, "Contextual dependency and lexical sets" (1996) y *Lexical analysis. Norms and exploitations* (2013); Blumenthal, "Profil combinatoire des noms: synonymie distinctive et analyse contrastive" (2002); Stefanowitsch & Gries, "Collostructions: Investigating the interaction of words and constructions" (2003) y "Corpora and grammar" (2009)— y proyectos lingüísticos —BDS, ADESSE, PDEV, *Sketch Engine*, el método *collostructional analysis*— vinculados con el concepto de *perfil combinatorio*, noción que han empleado como objeto de

estudio y como eje metodológico para la descripción lexicográfica y de otros fenómenos lingüísticos como las relaciones léxicas, el contraste interlingüístico, la fraseología o la traducción. La autora muestra claramente el desarrollo de este concepto, que se ha reconfigurado y enriquecido a medida que avanzan los estudios de corte empírico.

A lo largo del capítulo, rescata diferentes descripciones y aplicaciones de este *perfil combinatorio*, planteadas, todas ellas, en torno a la idea de la coocurrencia; por ejemplo, *a*) el patrón constructivo de un lema o una unidad lingüística cualquiera; *b*) su combinatoria léxica y sintáctica; *c*) su esquema de complementación; *d*) su perfil de uso normal, típico y frecuente; *e*) sus asociaciones preferidas, y, del estudio de García-Miguel, "El perfil combinatorio de los verbos en ADESSE: polisemia y parasinonimia de verbos de competición" (2014), *f*) el conjunto de probabilidades de coocurrencia con elementos léxicos y gramaticales. En conjunto, este inventario de definiciones, diversas pero complementarias, revela el interés metodológico que conlleva la noción, a saber, el de representar esquemáticamente el valor potencial de una unidad lingüística y, así, distinguirla de otras. El texto de Mas evidencia la indudable necesidad y el beneficio de realizar estudios de corpus representativos —es decir, de datos de lengua en uso— para comprender las relaciones entre la acepción de un lema y las construcciones sintácticas de las que éste participa, por medio del análisis puntual de concordancias que demuestren la frecuencia de coocurrencias significativas, ya sean éstas léxicas, sintácticas, semánticas o discursivas.

Olbertz, una de las editoras de este volumen, está a cargo del quinto capítulo, "Funciones pragmáticas en el portugués brasileño: un enfoque discursivo-funcional" (pp. 147-178). En este texto, indaga sobre el desarrollo, origen y funcionamiento de una marca formal de tópico en el portugués de Brasil (PB), y compara este fenómeno con la expresión y comportamiento de tópico y foco en la variedad de Portugal (PE) y en el español peninsular. Para ello, analiza muestras de uso registrado en corpus orales: *Iboruna* (PB), C-ORAL-ROM y *Português Falado* (PE) y PRESEEA (español, AdH). Tras una iluminadora presentación del marco discursivo-funcional y de las definiciones operativas empleadas en el trabajo —de *tópico*, *foco*, *función pragmática* y *función retórica*—, expone el uso brasileño del pronombre personal tónico pospuesto y adyacente a un sujeto frasal como marca de tópico de 3sg., como en el caso de *ele* en *meu filho ele costumava conversar* 'mi hijo solía conversar' (p. 157). La autora afirma que este uso pragmático del pronombre se originó a partir de un reajuste en

la expresión de sujeto dentro del paradigma verbal del PB, que disparó un proceso diacrónico hacia la obligatoriedad del sujeto pronominal, aún no cumplido.

Siempre basado en datos de corpus, el texto muestra las precondiciones de esta incipiente pragmaticalización de sujeto pronominal generalizado en marca de tópico (nuevo, simple o contrastivo): que, en una lengua de concordancia contextual, o *pro-drop*, su presencia es innecesaria sintácticamente, y que está empobrecido semánticamente, porque puede tener incluso un correferente inanimado (*coloco carne moída… até ela mudar de cor* 'coloco la carne molida hasta que cambie de color'). Los diferentes corpus, por último, evidencian que este fenómeno es exclusivo del PB. Olbertz entrega, sin duda, un texto denso y muy sugerente desde el punto de vista teórico y de la descripción de lenguas romances, que explota en profundidad los datos lingüísticos de uso registrados en los corpus consultados.

Domínguez, López y Barcala son los autores del sexto capítulo, que abre la segunda parte del volumen, dedicada al diseño y desarrollo de corpus. Este texto, "Corpus de Referencia do Galego Actual (CORGA): composición, codificación, etiquetaxe e explotación" (pp. 179-218) estudia la composición interna de este corpus del Centro Ramón Piñeiro para la investigación en humanidades (http://corpus.cirp.gal/corga). Los autores describen los grandes parámetros de selección de los materiales textuales del CORGA, a saber, *datación* (desde 1975 hasta la fecha), *tipo de documento* (v.gr. ficción, prensa, ensayo y guión) y *tema* (política, cultura y ciencias, entre otros). Con cerca de 40 millones de palabras ortográficas —que incluyen una muestra de oralidad de 25 horas de grabaciones radiofónicas, con transcripción ortográfica convencional no estandarizada—, CORGA logra representar cabalmente el gallego actual, inclusive en cuanto a su notoria variación ortográfica y morfológica, vinculada, sobre todo, a las diferentes propuestas de normativización por las que ha pasado esta lengua (1982, 1995, 2003).

En cuanto al proceso de elaboración, el capítulo pormenoriza las marcas que identifican fenómenos lingüísticamente relevantes (v.gr. *non normativo*, *inintelixible* y *alongamiento*); la estructuración en partes de los textos, para orientar finamente las consultas, y las características del lematizado y etiquetado gramatical, elaborado con XIADA, un sistema estadístico automático diseñado para este fin específico. Sobre esta tarea de anotación, sobresale la descripción del lexicón y del tratamiento de algunos ítems concretos, como el de los lemas simples frente a las formas amalgamadas (contracciones y verbos clitizados), entre muchos otros. Para cerrar, se presenta

la poderosa e intuitiva aplicación de recuperación de datos, con la que CORGA termina de demostrar su rendimiento, incontable e imprescindible, para los estudios léxicos y gramaticales de la lengua gallega actual.

En el séptimo capítulo, "CORILGA: un corpus para o estudo da variación e do cambio lingüístico no galego falado" (pp. 219-241), Fernández Rei y Regueira presentan el *Corpus oral informatizado da lingua galega* (CORILGA), del Instituto da Lingua Galega (USC) y de la Universidade de Vigo. Por sus características, éste es un recurso que se antoja fundamental para el desarrollo de tecnologías del habla y, sobre todo, para el estudio científico y empírico de la variación (diastrática y diafásica) y del cambio diacrónico (en tiempo aparente o real) del gallego oral, pues está conformado por muestras representativas de lengua vigente. Muchos de estos materiales fueron recogidos desde 1965, hasta la actualidad, para emplearse en diversos proyectos —entre los que destaca el *Atlas lingüístico galego*— y para las denominadas *teses de falas* e investigaciones sobre variación en lengua oral y patrimonio lingüístico.

CORILGA, al momento, cuenta con 105 horas de grabación que compilan una amplia variedad textual en cuanto a temática, registro y características de los informantes. Las grabaciones se complementan con una nutrida anotación linear que incluye transcripciones alineadas (ortográfica y fonética), información gramatical y datos sobre la consulta (como lema, lengua, y tema y tipología textual). Para su gestión, CORILGA cuenta con una base de datos y un sistema de búsqueda que permite filtrados complejos y obtención de concordancias para escuchar, leer y descargar en diferentes formatos (Excel, Elan, Praat). Por estas propiedades y por las utilidades que incorpora —de alineamiento *texto-voz*, reconocimiento de habla, de transcripción fonética automática, de lematización y de etiquetado—, este corpus, sin duda, resultará irremplazable y esencial en el estudio variacionista del gallego oral, en sus vertientes teórica y aplicada.

El octavo capítulo es la colaboración de Domínguez, Rivas, Santalla y Villapol, "Problemas afrontados en la etiquetación morfosintáctica del corpus ESLORA" (pp. 243-271). El texto reporta el proceso de anotación gramatical del *Corpus para el estudio del español oral*, cuyos materiales constan de entrevistas y conversaciones en la variedad gallega. Después de una compendiosa exposición de los rasgos de etiquetado de los corpus Val.Es.Co., COSER, C-ORAL-ROM y CORPES XXI, las autoras describen las fases y recursos del proceso de anotación diseñado para ESLORA: la adaptación del etiquetador XIADA, originalmente para el análisis de gallego; el etiquetario, compuesto

por 453 etiquetas que asignan valores o *hipervalores* según el contexto morfosintáctico; el diccionario, con cerca de 90 400 lemas, y el corpus de entrenamiento cero, conformado por muestras aleatorias de enunciados, en su mayoría extraídos también de ESLORA, con etiquetas automáticas revisadas y corregidas por los anotadores.

El capítulo continúa con un muy provechoso informe de numerosas dificultades de anotación ya identificadas en esta versión del corpus —y, en su gran mayoría, en vías de solucionarse— derivadas de la variedad oral representada. Entre éstas, todas cabalmente estudiadas, destaca el tratamiento de las marcas de oralidad (v.gr. *silencio*, *énfasis* o uso de otra *lengua*), de la transcripción de interjecciones no lingüísticas (*mmm, hmmm, mh…*), y de los problemas concernientes o a marcadores discursivos o a usos orales rutinarios no usuales en lengua escrita (*madre mía*, *qué va*, *tal*, *dijistes*). Las autoras entregan, en definitiva, esta utilísima memoria de problemas y propuestas de solución, coherentes con otros trabajos, a la que podrán recurrir futuros proyectos de lingüística de corpus como material de referencia.

Palacios, Barcala y Rojo entregan en el noveno capítulo, "El *Corpus de aprendices de español* (CAES) y sus aplicaciones para la enseñanza/ aprendizaje del español como lengua extranjera" (pp. 273-301), las características del CAES (http://galvan.usc.es/caes) y ejemplos muy valiosos de sus formas de explotación. El primer apartado informa sobre el diseño del corpus —a cargo del Instituto Cervantes y la Universidade de Santiago de Compostela—, recurso que compila y organiza muestras producidas por aprendices de español como L2, de niveles desde A1 hasta C1, hablantes nativos de seis distintas lenguas (árabe, mandarín, francés, inglés, portugués o ruso). Los materiales que conforman el corpus son textos escritos por estudiantes en sesiones para este fin específico. Los elementos lingüísticos obtenidos a partir de estos textos fueron lematizados y etiquetados, en procesos que, como un gran acierto, fueron desambiguados y corregidos después manualmente. Su aplicación de consulta, además, permite combinar filtros como L1, nivel, edad, etc., obtener estadísticas y descargar resultados.

La segunda parte del capítulo está enteramente dedicada a una exposición puntual de formas de usar el CAES. Como ejemplos de usos directos, los autores proponen, entre otros, la investigación en aprendizaje del español como L2, el análisis contrastivo de interlengua y la elaboración de materiales didácticos, como gramáticas, diccionarios y libros de texto. Como usos indirectos, el CAES se presenta como una herramienta invaluable para incidir en la formación

ISSN 0185-0121; e-ISSN 2448-6558; DOI: 10.24201/nrfh.v70i1.3792

docente, en la mejora de diseños curriculares y en la confección de instrumentos de evaluación. El gran aporte de este capítulo consiste, pues, en cómo los autores ejemplifican formas concretas para aprovechar al máximo un recurso de diseño tan cuidado como es el caso del CAES.

En el capítulo final, "Multifuncionalidad de los corpus paralelos, ejemplificada con el corpus alemán/ español PaGeS" (pp. 303-320), Doval y Jiménez presentan detalladamente las características distintivas de este corpus (www.corpuspages.eu), su contenido textual y partes nuclear y complementaria. Después de pasar revista a los corpus alemán/ español ya disponibles, los autores exponen los criterios con que se diseñó el PaGeS, que lo vuelven un recurso multifuncional ideal, tanto para investigadores en lingüística contrastiva, traducción, lexicografía y enseñanza de lenguas, como para estudiantes no especialistas. En cuanto a su composición, esta herramienta ofrece materiales en ambas lenguas, con sus respectivas traducciones alineadas, publicados desde 1960 por editoriales que respaldan su calidad. Éstos abarcan una variedad importante de géneros —narrativa, ficción infantil, ensayo y divulgación—, y de variantes dialectales —incluyen español americano, y alemán suizo y austríaco—, todo lo cual garantiza la variedad léxica y morfosintáctica indispensable para una correcta documentación y un procesado preciso de las lenguas fuente y meta.

Con respecto a este procesado, PaGeS alinea los textos, primero automática y luego manualmente, y los segmenta. Este método de preparación organiza los materiales en *bitextos* conformados por *bisegmentos*, es decir, por correspondencias oracionales, lo más paralelas y simétricas posible. Destaca, en este punto, la construcción de un diccionario automático bilingüe. A propósito de los medios para observar los resultados, sobresalen los tres tipos de búsquedas multinivel —simple, avanzada y formal— con que PaGeS atiende niveles distintos de complejidad de búsqueda, según las necesidades de consulta. Los autores entregan, en suma, una presentación muy minuciosa de los criterios en los que descansa esta herramienta, elaborada en la Universidade de Santiago de Compostela, y que resultan aportaciones metodológicas y disciplinarias invaluables para la lingüística general de corpus paralelos.

Resta únicamente señalar dos últimos asuntos. En primer lugar, que las editoras y autores entregan una obra colectiva cuidada, rigurosa y diversa que constituye un aporte muy significativo no sólo a la lingüística de corpus aplicada al español, gallego, inglés y alemán, sino también a la descripción de fenómenos concretos de lengua

española y portuguesa, sin mencionar las importantes propuestas teóricas que incluye. En segundo lugar, que este volumen es, en sí mismo, un exhaustivo repertorio de corpus que pone de manifiesto los enormes avances que ha tenido la disciplina, en cuanto a mejoras de diseño, accesibilidad, explotación y aplicación y, muy especialmente, a la creciente diversidad tipológica de estos recursos en el ámbito hispánico, representada aquí por la exposición de corpus generales, de aprendices, bilingües, paralelos, de entrenamiento e históricos.

Nueva Revista de Filología Hispánica (*NRFH*), LXX, 2022, núm. 1, 323-331
ISSN 0185-0121; e-ISSN 2448-6558; DOI: 10.24201/nrfh.v70i1.3792