PRESENTACIÓN

Los trabajos reunidos en *Corpus y construcciones. Perspectivas hispánicas* derivan de las contribuciones de sus autores al encuentro científico del mismo nombre celebrado en la Facultade de Filoloxía de la Universidade de Santiago de Compostela los días 22 y 23 de noviembre de 2018. El evento, organizado por el grupo de investigación Gramática del español, sirvió como marco para presentar un conjunto de investigaciones novedosas relacionadas con la lingüística de corpus que impulsaron un estimulante debate entre los participantes. Prueba del interés del encuentro y de la calidad de los trabajos expuestos es que una selección de diez aportaciones, en versión escrita ampliada y enriquecida, se compartan ahora de forma permanente con la comunidad académica a través de la edición de este anexo de la revista *Verba*.

El volumen integra trabajos relacionados con tres líneas preferentes de investigación del grupo Gramática del español desde su constitución: el análisis tanto sincrónico como diacrónico de estructuras gramaticales, las relaciones entre gramática y léxico, y la elaboración de corpus y bases de datos lingüísticas. Una fructífera combinación de estas tres orientaciones dio lugar en los años noventa del siglo XX a la construcción de la Base de datos sintácticos del español www.bds.usc.es, un recurso electrónico pionero elaborado a partir del análisis de un corpus de textos orales y escritos de las dos décadas anteriores (corpus ARTHUS: www.bds.usc.es/corpus.html), enriquecido posteriormente con información semántica y léxica en el proyecto ADESSE en la Universidade de Vigo http://adesse.uvigo.es/.

Desde el núcleo inicial de estudios centrados en las estructuras sintácticas, las investigaciones de los miembros del grupo se han ampliado con el análisis de nuevos problemas y la adopción de nuevos enfoques. Cabe destacar el creciente interés por los fenómenos de contacto, por un lado, y por el contraste entre lenguas y variedades, por otro. En ambas perspectivas se atribuye una relevancia crucial a los datos de uso como fundamento empírico de la investigación.

La importancia concedida al uso lingüístico real, oral y escrito, como objeto imprescindible de una aproximación funcional al estudio de la lengua no solo ha estimulado las investigaciones basadas en materiales auténticos, sino que ha evidenciado la necesidad de disponer de corpus que proporcionen la información requerida en las condiciones de accesibilidad adecuadas. En los últimos años, el trabajo del grupo sobre corpus se ha ampliado con la creación de nuevos recursos y nuevas herramientas de tratamiento y anotación de textos que facilitan su explotación tanto en lingüística teórica y descriptiva como en el campo de la lingüística aplicada.

CONTENIDO DEL VOLUMEN

El volumen tiene dos partes, que se relacionan respectivamente con los dos ámbitos de desarrollo de la lingüística de corpus; por un lado, con el análisis de fenómenos lingüísticos basados en datos extraídos de corpus y, por otro, con el diseño, elaboración y enriquecimiento de corpus con vistas a una adecuada recuperación y explotación de la información que contienen.

En cuanto a las lenguas sobre las que versan los trabajos, aunque el español es objeto de estudio de buena parte de ellos, el volumen incluye aportaciones sobre el gallego (capítulos 6 y 7), el inglés (capítulo 1), el portugués (capítulo 5) y el alemán (capítulo 10). Además, varios capítulos muestran la necesidad de un enfoque plurilingüe, bien para dar cuenta de fenómenos de variación y cambio en situaciones de contacto, como ocurre con el español y el inglés en Nuevo México (capítulo 1), bien para desarrollar recursos lingüísticos para la enseñanza de lenguas extranjeras o para la traducción, como el corpus de aprendices CAES (capítulo 9) o el corpus paralelo alemán/español PaGeS (capítulo 10).

La primera parte, dedicada a los estudios gramaticales con datos de corpus, se abre con el capítulo de Rena Torres Cacoullos y Catherine Travis «Gramáticas en contacto en un corpus bilingüe», cuyo objetivo es revisar la validez de la hipótesis de la convergencia gramatical, es decir, determinar si el cambio lingüístico se debe realmente al contacto y si el estatus minoritario o dominante de cada lengua determina el sentido del cambio. El estudio se fundamenta en la observación sistemática de las muestras de habla espontánea que integran el *Corpus bilingüe español-inglés de Nuevo México*, que por su configuración constituye un recurso idóneo para comparar el uso que hacen de ambas lenguas los hablantes bilingües de la comunidad objeto de estudio. Mediante una metodología cuantitativa cuidadosamente diseñada, que confronta los datos del corpus bilingüe con la información obtenida de dos corpus monolingües de español e inglés, se establece el grado de semejanza

entre las gramáticas bilingües de inglés y español y entre las correspondientes gramáticas monolingües en tres tipos de estructuras variables: las perífrasis progresivas, el uso de indicativo y subjuntivo en cláusulas subordinadas sustantivas y la expresión variable del sujeto pronominal. Los resultados del estudio contradicen las propuestas anteriores de convergencia gramatical y sustentan la continuidad de la independencia de las dos gramáticas de los hablantes bilingües.

El segundo capítulo, elaborado por Anton Granvik y titulado «Sobre los orígenes de la construcción encapsuladora en español», se basa en los datos del corpus del *Nuevo diccionario histórico del español* para ofrecer un detallado análisis diacrónico de la construcción por la que ciertos sustantivos abstractos como *causa*, *condición* o *idea* remiten a una unidad de información compleja de tipo proposicional. Para superar las dificultades que plantea el criterio de la identidad experiencial en la determinación de la función encapsuladora de cada sustantivo, el estudio parte de una interpretación formal y esquemática de la construcción, compatible con el enfoque construccionista, e incorpora tres propiedades gramaticales como criterios operacionales de encapsulación: la determinación de la frase nominal, su función sintáctica y el tipo de unidad introductora. Esta aproximación metodológica permite al autor establecer una escala de tipicidad a partir de una amplia muestra de usos de nueve sustantivos entre los siglos XIII y XX. El estudio se completa con un minucioso análisis cualitativo semántico-cognitivo y textual de los elementos seleccionados que permite detectar diferencias funcionales condicionadas por la semántica léxica propia de cada sustantivo.

La gramática de construcciones y la lingüística de corpus constituyen el marco teórico y metodológico del capítulo 3, «*Entre miradas de asombro*: aportaciones de la Lingüística de Corpus al estudio de una construcción con la preposición *entre*», de Belén López Meirama y Carmen Mellado Blanco. En este caso la base empírica del análisis se extrae del CORPES XXI a partir de una búsqueda inicial de proximidad a la que se aplica un filtrado manual para seleccionar la combinación [*entre* + sustantivo$_{\text{plural/corporal}}$]. El estudio detallado de todas las secuencias que presentan tal estructura abarca tanto los aspectos morfológicos y sintácticos como sus valores semánticos y pragmáticos, con especial atención a la unidad léxica variable de la construcción, identificada como un sustantivo, plural o en coordinación, de comunicación o expresión corporal. El análisis revela el predominio de una configuración prototípica de la unidad fraseológica en torno a un número reducido de sustantivos nucleares que dan cuenta del 50% de los casos. Se observa asimismo la existencia de un efecto de coerción semántica ejercido por el primer sustantivo coordinado

sobre el segundo. Como resultado de alcance más general, el trabajo ofrece el diseño y la aplicación de una propuesta metodológica extensible a otras unidades fraseológicas preposicionales.

En el capítulo 4, «En torno al concepto de *perfil combinatorio*», Inmaculada Mas Álvarez realiza un recorrido por diferentes propuestas que han ido configurando la noción de 'perfil combinatorio' como una aportación clave para el enriquecimiento de las descripciones lexicográficas a partir de los resultados de la lingüística de corpus. El concepto se fundamenta en que los datos de frecuencia y coocurrencia léxico-gramatical obtenidos a través de concordancias permiten identificar, con base en el uso real, los patrones constructivos de las unidades analizadas. En el texto se describen las características de algunos recursos que incorporan de manera sistemática información sobre la combinatoria sintáctica verbal, como la BDS y ADESSE, este último incluyendo propiedades semánticas y definiciones léxicas. Se informa asimismo sobre las opciones que ofrece la herramienta *Sketch Engine* para analizar en detalle el perfil combinatorio de los elementos léxicos y establecer comparaciones entre perfiles; y finalmente se resume el método del *collostructional analysis*, que mide el grado de atracción entre unidades léxicas y construcciones y ha demostrado su utilidad en el análisis de diversas relaciones léxicas en diferentes lenguas.

Cierra la primera parte del volumen el capítulo de Hella Olbertz «Funciones pragmáticas en el portugués brasileño: un enfoque discursivo funcional». La autora parte de la comparación entre el español y el portugués para examinar con detalle la expresión de las funciones pragmáticas de tópico y foco en el portugués brasileño. El estudio se basa en un detallado análisis cualitativo del uso registrado en corpus orales comparables: PRESEEA de Alcalá de Henares para el español, *Iboruna* para el portugués del Brasil y C-ORAL-ROM y una parte de *Português Falado* para la variedad de Portugal. Los conceptos funcionales empleados en el análisis —tópico y foco, agente y paciente, sujeto y objeto— se definen y contextualizan en el modelo de la gramática discursivo-funcional, del que se ofrece una breve pero ilustrativa presentación. En el núcleo del trabajo se explican de forma pormenorizada y empíricamente fundamentada (i) los cambios que ha experimentado el portugués brasileño en la expresión personal del sujeto y (ii) cómo la progresiva sobrecarga funcional de la concordancia verbal de 3ª persona de singular provocó la generalización del pronombre sujeto, cuya desemantización y pragmaticalización ha dado origen a una marca gramatical de la función de tópico. Se establece así un contraste entre el español, lengua con una marca propia de la función focal, y el portugués de Brasil, que ha desarrollado un mecanismo innovador para identificar el tópico.

La segunda parte del volumen integra cinco capítulos centrados en el diseño y desarrollo de corpus. En el primero de ellos, el capítulo 6, Eva María Domínguez Noya, María Sol López Martínez y Francisco Mario Barcala Rodríguez presentan «Corpus de Referencia do Galego Actual (CORGA): composición, codificación, etiquetaxe e explotación». El trabajo informa en primer lugar de la composición y estructuración interna del corpus, que abarca desde 1975 hasta la actualidad y alcanza en la versión 3.1 una extensión de 40 178 271 palabras. Los materiales del CORGA consisten en una amplia variedad de textos escritos y en una muestra oral de 25 horas de emisiones radiofónicas transcritas y alineadas con el audio. Se destaca la importancia del diseño del corpus, tanto en lo que atañe a los criterios de clasificación textual (fecha, tipo o género, área temática) como en lo referente al tratamiento de la variación gráfica —que se resuelve introduciendo la categoría de 'hiperlema'— y morfológica. En el capítulo se expone el protocolo que siguen los documentos en el proceso de construcción del corpus y se presenta el sistema de etiquetación morfosintáctica llevado a cabo con el etiquetador del gallego actual XIADA, desarrollado en relación con el corpus CORGA y adaptado a sus necesidades. El trabajo se completa con la descripción del sistema de recuperación de los datos del corpus a través de la aplicación de consulta.

A continuación, en el capítulo titulado «CORILGA: un corpus para estudar a variación e o cambio do galego falado», Elisa Fernández Rei y Xosé Luís Regueira contextualizan la creación de un corpus oral actual en el marco de las iniciativas impulsadas por el *Instituto da Lingua Galega* (ILG) a lo largo de las últimas décadas. Buena parte de las 105 horas de grabación que recoge el corpus proceden de proyectos anteriores y forman un valioso conjunto de materiales de habla que se ponen ahora a disposición pública en las condiciones de acceso y consulta adecuadas para facilitar el estudio de la variación diastrática, diafásica e incluso diacrónica, ya que los registros sonoros se extienden desde 1965 hasta el momento actual. Para la transcripción y anotación del corpus en diferentes niveles se emplea el programa ELAN, que permite integrar diferentes recursos de tecnología del habla desarrollados en colaboración con el Grupo de Tecnoloxías Multimedia de la Universidade de Vigo, entre las que destacan las herramientas de alineación texto-voz, reconocimiento de voz y transcripción automática.

El capítulo 8, elaborado por Eva M.ª Domínguez Noya, Raquel Rivas Cabanelas, M.ª Paula Santalla del Río y Rebeca Villapol Baltar, trata de «Problemas afrontados en la etiquetación morfosintáctica del corpus ESLORA». Tras resumir las condiciones de etiquetación de varios corpus orales desarrollados con anterioridad, el trabajo ofrece información sobre los recursos

utilizados en el tratamiento de ESLORA, un corpus que recoge entrevistas y conversaciones de hablantes de español en Galicia. Se presentan las características del etiquetador, el etiquetario, el diccionario y el corpus de entrenamiento, y se describen las fases del proceso de etiquetación. El núcleo del capítulo aborda algunas de las dificultades que se plantearon en el proceso de revisión manual de los resultados de la etiquetación automática junto con las soluciones adoptadas en cada caso. Se explican, entre otras, las opciones elegidas para la anotación de los numerales, los ruidos comunicativos, los conectores discursivos, el tratamiento de formas que se apartan del estándar normativo y ciertos usos de formas frecuentes como *que* y *tal*. Como conclusión, el trabajo subraya la necesidad de adaptar los recursos y las decisiones de anotación a las particularidades de los materiales anotados, en este caso a las características específicas de un corpus oral como ESLORA.

«El Corpus de Aprendices de Español (CAES) y sus aplicaciones para la enseñanza/aprendizaje del español como lengua extranjera» es el título del capítulo 9, en el que Ignacio Palacios Martínez, Francisco Mario Barcala Rodríguez y Guillermo Rojo describen las características del corpus, las fases de su construcción y etiquetación e ilustran con casos prácticos las principales líneas de explotación de la información que contiene. En su versión 1.2 de agosto de 2018 el CAES comprende cerca de 600 000 elementos lingüísticos correspondientes a la producción escrita de una amplia muestra de aprendientes de español con niveles desde A1 a C1 hablantes iniciales de alguna de las siguientes lenguas: árabe, chino mandarín, francés, inglés, portugués y ruso. El hecho de que la anotación morfosintáctica de los textos haya sido rigurosamente desambiguada de forma manual incrementa notablemente la utilidad del recurso, tanto como fuente de información para incidir de forma efectiva en el proceso de aprendizaje, como por las posibilidades de aprovechamiento de los datos en la elaboración de actividades de aula y materiales didácticos. El trabajo muestra asimismo que el corpus, por sus características y por las facilidades de consulta que ofrece, constituye un recurso valioso en la formación del profesorado de ELE y como fuente de referencia para el diseño curricular.

En el capítulo final del volumen, que lleva por título «Multifuncionalidad de los corpus paralelos, ejemplificada con el corpus alemán / español PaGeS», Irene Doval y Tomás Jiménez dan cuenta de la composición del corpus PaGeS, del proceso de compilación y de las diferentes opciones de recuperación de datos que ofrece. PaGeS es un corpus formado en su parte nuclear por textos fundamentalmente narrativos alemanes y españoles escritos en las últimas décadas y alineados con sus traducciones al español y al

alemán. La versión 2.0 de abril de 2019 contiene 28 millones de unidades distribuidas de forma equilibrada entre ambas lenguas. En el capítulo se describen los pormenores de la preparación de los materiales y las dificultades de segmentación y alineado de los textos junto con las soluciones alcanzadas, lo que implica, por una parte, la selección de herramientas computacionales adecuadas, y por otra, un laborioso trabajo de validación manual. La presentación se completa con la ilustración de las posibilidades que ofrece el motor de búsqueda para recuperar la información contenida en el corpus. Se apuntan finalmente las previsiones de enriquecimiento y mejora del recurso en un futuro inmediato, como la ampliación del alineado al nivel de la palabra y la etiquetación morfosintáctica de los textos.

AGRADECIMIENTOS

Tanto la edición del presente volumen como la organización del encuentro en el que se presentaron versiones previas de los trabajos aquí incluidos se enmarcan en la estrategia de consolidación propuesta por el grupo de investigación Gramática del español para el trienio 2017-2019. En este período el grupo fue beneficiario de una ayuda del Programa de consolidación y estructuración de unidades de investigación competitivas, en la modalidad de Grupos con potencial de crecimiento, concedida por la Consellería de Cultura, Educación e Ordenación Universitaria y la Consellería de Economía, Emprego e Industria de la Xunta de Galicia a través de la Axencia Galega de Innovación (ref. nº ED431B 2017/39). Para la organización del encuentro contamos asimismo con la financiación del proyecto de investigación ESLORA+ (ref. FFI2017-86379-P) subvencionado por la Agencia Estatal de Investigación (AEI) y por el Fondo Europeo de Desarrollo Regional (FEDER).

Por último, deseamos agradecer públicamente la colaboración generosa de las personas que elaboraron los dictámenes de los trabajos enviados para publicación en el volumen. Su contribución ha sido imprescindible para garantizar la calidad de los diez capítulos finalmente seleccionados, cuyas versiones definitivas se han beneficiado de los pertinentes comentarios aportados en los informes de evaluación.

Marta Blanco, Hella Olbertz, Victoria Vázquez Rozas