# INTRODUCTION

## Fraseología computacional y de corpus aplicadas al español

Este número monográfico de *Romanica Olomucensia* sobre fraseología computacional y de corpus aplicadas al español nace como *publicación derivada* de la Conferencia Internacional EUROPHRAS 2022 «Computational and Corpus-based Phraseology», celebrada del 28 al 30 de septiembre de 2022 en la Universidad de Málaga bajo la dirección de Gloria Corpas Pastor (Universidad de Málaga) y Ruslan Mitkov (Universidad de Wolverhampton). La publicación no estaba en ningún caso pensada de forma exclusiva para los participantes del congreso, sino que se realizó una convocatoria abierta a todos los especialistas en el tema. Su propósito no era otro que visibilizar y presentar internacionalmente los diversos avances en el campo de la fraseología basada en corpus informatizados y el procesamiento computacional de unidades fraseológicas (UF) en español, así como sus aplicaciones prácticas en los campos de la traducción, la interpretación, la lexicografía y el proceso de enseñanza/aprendizaje del español como lengua extranjera. Este número monográfico de *Romanica Olomucensia* aspira, además, a convertirse en una suerte de continuación del 32.1, publicado en 2020 y centrado en una de las aproximaciones más productivas en la investigación lingüística actual: la Gramática de Construcciones aplicada a los estudios de fraseología (*cfr.* Mellado Blanco – Gutiérrez Rubio 2020).

La fraseología como disciplina lingüística en España vivió un primer momento de auge en los años noventa del siglo pasado gracias a la aparición, en un periodo relativamente breve de tiempo, de un importante número de trabajos a cargo de algunas de sus principales figuras, entre los que cabría destacar muy especialmente, por haberse convertido en obras de referencia generalizada, *Manual de fraseología española* de Gloria Corpas Pastor (1996) y *Aspectos de fraseología teórica española* de Leonor Ruiz Gurillo (1997). Además de estos trabajos de carácter teórico, que trataban de poner un poco de orden en el mar de confusión en el que vivía la fraseología por aquellos tiempos, es digna de mención la publicación de los primeros diccionarios fraseológicos monolingües basados en rigurosos criterios fraseográficos, entre los que el *Diccionario fraseológico del español moderno* de Varela y Kubarth (1994) ocupa un lugar privilegiado. Por otra parte, en aquel tiempo las obras de fraseología aplicada a la traducción y la interpretación eran prácticamente desconocidas, y las centradas en

el proceso de enseñanza/aprendizaje, aún muy escasas, si bien *La enseñanza de las unidades fraseológicas* de Inmaculada Penadés Martínez (1999) puede considerarse un trabajo pionero en este campo. Hemos hablado de *momento de auge* y, sin embargo, quizá sería más correcto hablar de *momento de eclosión,* ya que hasta entonces eran escasísimos los trabajos científicos sobre fraseología realizados en España más allá de aquella flor en el desierto que supusieron los capítulos de *Introducción a la lexicografía moderna* que Julio Casares (1950) dedicó a la fraseología (cuya traducción al ruso de 1958 sigue siendo, curiosamente, la única obra de este autor catalogada en la biblioteca de la Universidad Palacký). Los principales trabajos sobre fraseología española publicados entre la obra seminal de Casares y los años noventa salieron de las plumas de especialistas hispanoamericanos: el venezolano afincado en Alemania Alberto Zuluaga, entre cuyas publicaciones destaca su *Introducción al estudio de las expresiones fijas* (1980), y las cubanas Zoila Carneado Moré y Antonia María Tristá Pérez, que, en el marco de la asentada tradición fraseológica soviética, publicarían *Estudios de fraseología* (1985), entre otras obras de gran interés para su momento.

Los aproximadamente treinta años transcurridos desde ese primer auge de la fraseología española han sido enormemente fructíferos. El panorama es hoy en día completamente distinto, casi irreconocible: la fraseología, considerada ya de forma unánime una disciplina lingüística autónoma, cuenta en España con multitud de especialistas, algunos de reconocido prestigio internacional, que han desarrollado enormemente este campo del saber tanto desde la perspectiva teórica como aplicada. En este sentido, resulta necesario destacar dos obras fraseográficas que recogen, por primera vez, un número muy importante de unidades fraseológicas compiladas, además, con gran rigurosidad: la segunda edición del *Diccionario fraseológico documentado del español actual: locuciones y modismos españoles*, a cargo de Seco, Andrés y Ramos (2017), y el *Diccionario de locuciones idiomáticas del español actual* en cuya versión completa aún trabaja Penadés Martínez (2019).

El desarrollo continuo durante tres décadas de la fraseología como disciplina lingüística en las universidades y centros de investigación españoles, en conjunción con los significativos avances en las tecnologías lingüísticas (TL), el procesamiento de lenguaje natural (PLN) o la inteligencia artificial (AI) logrados en los últimos años, ha posibilitado lo que podríamos denominar un *segundo momento de auge* de la fraseología en España. Los corpus de creación propia hechos a mano, en que se basó una buena parte de las obras teóricas de los años noventa, han dado paso a *gigacorpus* que ponen a disposición del investigador miles de millones de palabras accesibles a través de un *software* que posibilita, además, búsquedas mucho más precisas y ambiciosas que, por ejemplo, los clásicos corpus de la Real Academia. La posibilidad de contar con corpus informáticos de semejante tamaño resulta de especial importancia en el estudio fraseológico, ya que, aunque el uso de UF sea una constante tanto en el discurso oral como en el escrito –recordemos, a modo de ejemplo, que Erman y Warren (2000) afirman que al menos el 50 % del corpus que analizan consta de *frases preconstruidas*–, la frecuencia de cada una de las unidades individuales tiende a ser muy baja (Moon 1998; Corpas Pastor 2021), resultando excepcionales

aquellas UF que alcanzan una ocurrencia por millón de formas (Gutiérrez Rubio 2022). Por otra parte, centros tecnológicos como el Instituto Universitario de Investigación de Tecnologías Lingüísticas Multilingües (IUITLM)[1] de la Universidad de Málaga trabajan en el desarrollo de nuevas herramientas lingüísticas que faciliten la labor de traductores e intérpretes y que, en gran medida, incluyen entre sus innovadores proyectos el procesamiento de la fraseología. Además de mejorar las capacidades formativas y profesionales de traductores e intérpretes, que encuentran en las UF uno de sus peores enemigos, resulta fundamental realizar nuevas investigaciones basadas en corpus y otras tecnologías computacionales para lograr que la subjetividad y la intuición de los autores sean sustituidas por datos empíricos tanto en la compilación de futuros diccionarios fraseológicos (bilingües y multilingües), como en la creación de nuevos materiales para la enseñanza/aprendizaje de lenguas extranjeras y, muy especialmente, respecto a la selección de UF y su introducción por niveles de dominio.

El contexto actual de la investigación fraseológica en España está estrechamente ligado, pues, a los desarrollos recientes de las tecnologías lingüísticas en general y, más concretamente, a los avances de la fraseología computacional y la fraseología basada en corpus. Trabajos pioneros como los contenidos en los volúmenes editados por Monti *et al.* (2018), Corpas Pastor y Colson (2020) o Corpas Pastor y Mitkov (2019, 2022) han desbrozado el terreno que ahora empiezan a transitar los fraseólogos que se ocupan del español y de su traducción hacia y desde otras lenguas (*cfr.* Corpas Pastor *et al.* 2021).

Es precisamente en este contexto de innovación metodológica de la investigación fraseológica aplicada al español en el que se desarrollan los diez artículos que, tras superar un exigente proceso de selección, salen ahora publicados en este número 35.1 de *Romanica Olomucensia.* No queremos pasar a presentar brevemente estos diez trabajos sin antes expresar nuestro sincero agradecimiento a las decenas de especialistas en fraseología que han realizado las evaluaciones anónimas de los originales recibidos para este número: sin su generosa colaboración, esta publicación no habría sido posible.

Si hablamos de nuevos recursos en la investigación lingüística, en general, y fraseológica, en particular, uno destaca sobre los demás en los últimos años: los corpus de la familia TenTen y, muy especialmente, el corpus de español esTenTen18 accesible a través de Sketch Engine, un complejo sistema de compilación, procesamiento, análisis y gestión de corpus desarrollado en la ciudad de Brno, a apenas unas decenas de kilómetros de Olomouc. Comenzamos con la presentación de uno de los cuatro artículos de este número monográfico que han hecho uso de esta tecnología: el trabajo escrito a cuatro manos entre Antonio Pamies y José Manuel Pazos titulado «Frasemas combinados 'a pedir de boca'» (p. 139-156). En él, a través de la aplicación de criterios cuantitativos y cualitativos, se aporta nueva luz sobre la (inesperadamente) rica combinatoria de la locución adverbial que da nombre al artículo. El elevado número de verbos de diversas clases semánticas que, junto al canónico *salir,* se

[1] <https://iuitlm.uma.es/en/home/>.

emplean en combinación con *a pedir de boca* aporta «nuevos argumentos en favor de una concepción dinámica de la fraseología, cuya fijación puede ser a veces de carácter conceptual […] más que de carácter formal, como exigiría la fijación fraseológica en el sentido tradicionalmente asumido» (p. 150).

Otro artículo basado en datos extraídos del corpus esTenTen18 es «Fraseología de las emociones en colores: el arte de la comunicación e implicaciones pedagógicas» (p. 107-120) de Beatriz Martín-Gascón. Esta investigadora de la Universidad Complutense de Madrid presenta los resultados de un estudio de tres unidades fraseológicas de color-*emoción* en español: *ponerse rojo, ponerse blanco* y *ponerse negro*. Los datos obtenidos del corpus aclaran la frecuencia de uso de estas tres UF, así como la emoción a la que se asocian más comúnmente. Además, Martín-Gascón propone una serie de recomendaciones pedagógicas específicas que «proporcionan estrategias concretas para aplicar la metáfora y la metonimia en el aula de lengua extranjera» (p. 117).

En «Glozoo: un glosario trilingüe (ES/EN/DE) basado en corpus para la traducción de zoologismos manipulados» (p. 71-88), Carlos Manuel Hidalgo-Ternero y Marina Rueda-Martín presentan una innovadora metodología para la formación de traductores y, más concretamente, para afrontar con éxito una de las tareas más exigentes que se les puede presentar: la búsqueda de equivalentes de UF que han sufrido algún tipo de manipulación (o desautomatización) y que, además, no cuentan con un equivalente directo en la lengua meta. Para enfrentarse a este difícil reto, los autores han diseñado el glosario trilingüe (español, inglés, alemán) de zoologismos Glozoo –creado a partir de corpus paralelos y monolingües disponibles en Sketch Engine–, que ayuda al traductor a «crear un equivalente fraseológico *ad hoc,* cuya manipulación representaría asimismo una dilogía en la que ambos niveles [el figurado y el literal] estuvieran presentes de forma simultánea, al igual que la UF en el TO [texto origen]» (p. 84). El artículo incluye, además, una unidad didáctica a partir de un artículo publicado en *El País* que permite ejemplificar las aplicaciones prácticas en la formación de traductores a través de esta nueva herramienta.

Si Carlos Manuel Hidalgo-Ternero y Marina Rueda-Martín desarrollan nuevas herramientas enfocadas a la labor del traductor en formación, Mahmoud Gaber, en su trabajo «Cómo dominar la fraseología y automatizar el proceso de documentación: una solución tecnológica para la formación de intérpretes en la combinación español<>árabe» (p. 55-70), propone una tecnología innovadora con una finalidad semejante, si bien enfocada a la interpretación. Su investigación se centra en el análisis de patrones colocacionales documentados en corpus comparables en el ámbito de la ciberseguridad compilados de forma automática, nuevamente a través de Sketch Engine, a partir de materiales orales y escritos. La transcripción de los discursos se ha realizado mediante reconocimiento automático del habla, lo que facilita en gran medida el proceso de preparación del intérprete para un nuevo encargo profesional. La compilación y el tratamiento adecuados de estos corpus son un gran aliado para el intérprete a la hora de «sistematizar el proceso de documentación y extracción» (p. 68). Además, este tipo de análisis comparativo de colocaciones ofrece nuevos

datos respecto a la fraseología contrastiva árabe-español, una combinación insuficientemente desarrollada hasta el momento.

Un trabajo cuyo análisis se basa en un corpus creado *ad hoc* de forma más tradicional, si bien las UF extraídas se ponen en relación con los datos obtenidos de uno de los *gigacorpus* de Sketch Engine, es el que lleva por nombre «Las modificaciones fraseológicas en los titulares españoles» (p. 121-138). En él, Florentina M. Mena-Martínez y María F. Sáez Martínez, al igual que en el artículo ya presentado de Carlos Manuel Hidalgo-Ternero y Marina Rueda-Martín, abordan el tema de las modificaciones fraseológicas (o UF desautomatizadas). Su corpus –fruto de búsquedas manuales en seis revistas del corazón y de divulgación en línea– está formado por 286 titulares que incluyen unidades fraseológicas. El número de titulares con algún tipo de *alteración creativa* se sitúa en el 28,3 %, lo que demuestra la elevada frecuencia de este recurso. Nuevamente se ha empleado el corpus de español esTenTen18, si bien, en este caso, tan solo para determinar qué UF documentadas en los titulares de prensa pueden ser consideradas canónicas y cuáles responden a modificaciones no usuales. Entre las conclusiones más llamativas de la investigación destaca que, aunque se han documentado modificaciones en todos los tipos de unidades analizadas (colocaciones, locuciones, paremias, citas modernas y unidades fraseológicas pragmáticas), son las locuciones las que presentan un porcentaje relativo más elevado en este tipo de discurso, así como que «con respecto a la taxonomía de las modificaciones, también se han identificado ejemplos que ilustran todos los tipos, si bien [...] las modificaciones internas duplican en número a las externas» (p. 135).

A diferencia de los trabajos anteriores, que emplean, en mayor o menor medida, corpus accesibles a través de Sketch Engine, Belén López Meirama, en su trabajo titulado «*¡Con lo felices que éramos!* Otra mirada sobre la construcción [*con* ART (X) *que* V] del español» (p. 89-106), parte de los datos obtenidos de un corpus *tradicional* de la RAE, concretamente del Corpus del Español del Siglo XXI (CORPES). Su análisis, basado en un enfoque construccional de corte cognitivista, pretende ampliar significativamente la descripción de esta construcción más allá de sus peculiaridades formales y significado básico recogidos en las gramáticas. Esta especialista de la Universidad de Santiago de Compostela concluye que «la perspectiva construccional, que concibe la gramática como una red de construcciones interrelacionadas, propicia análisis más completos y también más coherentes» (p. 104).

M.ª Auxiliadora Castillo Carballo y Juan Manuel García Platero presentan, en «Neoformas fraseológicas a partir del corpus Banco de neologismos del Centro Virtual Cervantes y Antenario» (p. 9-21), un análisis de neologismos fraseológicos del español –entendiendo la fraseología desde una perspectiva *laxa* que incluiría compuestos sintagmáticos, colocaciones y locuciones– especialmente centrado en aquellos que aluden a la reciente pandemia de COVID-19. Entre las conclusiones del artículo destacaría que, frente «a las formas neológicas no pluriverbales, en las que se pueden constatar mecanismos lexicogenésicos que otorgan una voluntad expresiva, en las unidades fraseológicas de reciente creación se ha podido observar el predominio del carácter esencialmente designativo, lo que justifica la preferencia por la nominalidad» (p. 19).

En «La traducción e interpretación de nombres de organizaciones en el eurolecto inglés y español: un estudio basado en corpus» (p. 157-172), Fernando Sánchez Rodas emplea el sistema VIP (plataforma modular que incluye compilación semiautomática, gestión y consulta de corpus, entre otras funcionalidades) para crear dos corpus comparables inglés-español (uno escrito y otro oral) a partir de material del Parlamento Europeo. Además del empleo de una nueva tecnología para la investigación fraseológica, el artículo presenta otra novedad respecto a la fraseología española *tradicional:* no se centra en el análisis de las unidades fraseológicas comúnmente aceptadas –locuciones en la perspectiva estrecha; colocaciones, locuciones, paremias y fórmulas oracionales en la ancha–, sino en las *expresiones multipalabra* (EMP), lo que representa una forma menos estricta de entender los límites de la fraseología que, por otra parte, se halla bastante extendida más allá de las fronteras de España. En su estudio, este lingüista de la Universidad de Málaga subraya el potencial de las entidades nombradas (*nombres propios* en la terminología tradicional) multipalabra «como puerta de acceso [...] al entendimiento de que la traducción y la interpretación no son etiquetas absolutas impuestas *ex machina*, sino de que cada género textual (no)traducido o (no)interpretado se rige por su propia gramática construccional, repartiendo un número determinado de construcciones mediadoras y no mediadoras en función de parámetros como el idioma, las normas y eventos que rodean a la producción textual, etc.» (p. 171).

Miguel Da Corte y Jorge Baptista abordan, en un artículo que lleva por título «Etiquetaje de expresiones multipalabra en ensayos escritos por nativos y no nativos de español en un curso de desarrollo de gramática y composición» (p. 23-40), un análisis de las expresiones multipalabra documentadas en un corpus de ensayos escritos por estudiantes, tanto hablantes nativos como no nativos de español, de un curso de un instituto de educación superior estadounidense centrado en la gramática y composición escrita de textos en español. Los autores sometieron el corpus a tres experimentos distintos mediante el sistema de gestión de corpus Orange: a) corpus sin etiquetaje, b) corpus con etiquetaje y enlace de las EMP y c) corpus sin etiquetaje, pero con enlace. Los resultados obtenidos son un buen indicador de las «áreas de desarrollo léxico (y tipos de EMP) que deberían tomarse en cuenta para enriquecer las estrategias pedagógicas y de competencia lingüística en los programas de estudios orientados a la enseñanza del español como lengua extranjera» (p. 38).

Por último, en «La traducción automática de unidades fraseológicas usadas en *Manolito Gafotas* al árabe: entre la corrección lingüística y la aceptabilidad cultural» (p. 41-53), Mohamed El-Madkouri Maataoui y Beatriz Soto Aranda estudian las soluciones ofrecidas por las herramientas de traducción automática Google Translate y Reverso Context (en la combinación español-árabe) de cinco UF (una colocación, tres locuciones y una frase proverbial) extraídas de la popular obra de Elvira Lindo. Estas soluciones son comparadas con las traducciones de estas mismas UF ofrecidas por seis hablantes bilingües con formación en traducción. En sus conclusiones, los autores señalan que, aunque «los traductores automáticos neuronales tienden

a proponer soluciones válidas para la traducción de UFS cuando reciben suficiente *input* lingüístico, [...] en el caso de hablantes bilingües y biculturales, y con formación en traducción, sus soluciones son culturalmente más aceptables, recurriendo en muchos casos a UFS equivalentes a las españolas» (p. 52).

En suma, se trata de diez artículos que, aun presentando no pocas diferencias teórico-metodológicas y centrados en distintas unidades fraseológicas (e incluso en algunas combinaciones de palabras que, para una buena parte de los fraseólogos, se hallan más allá de los límites de esta disciplina) evidencian la gran relevancia que los corpus informáticos –y, entre ellos, muy especialmente, los de la familia TenTen accesibles a través de Sketch Engine– y las nuevas tecnologías de procesamiento lingüístico –como las herramientas de traducción automática, las funcionalidades de corpus del sistema VIP, las tecnologías de reconocimiento automático del habla o el sistema de minería de datos Orange– tienen actualmente en el análisis fraseológico. Sin querer restar importancia a trabajos de un corte más tradicional, que continúan aportando información fundamental sobre la naturaleza y las características de las UF, consideramos que estas y otras tecnologías abren la puerta a una nueva forma de entender la investigación fraseológica –especialmente en su variante aplicada–, disciplina lingüística esta que ha pasado, en apenas tres décadas, de la periferia (por no hablar directamente de marginalidad) a la centralidad de los estudios lingüísticos. Esperamos que este número monográfico sirva para consolidar y difundir estas nuevas aproximaciones al estudio de la fraseología, así como para favorecer su futuro desarrollo tanto en español, como en otras lenguas.

**Referencias bibliográficas**


Carneado Moré, Zoila – Tristá, Antonia María (1985), *Estudios de fraseología*, La Habana: Editorial de Ciencias Sociales.

Casares, Julio (1950), *Introducción a la lexicografía moderna*, Madrid: Consejo Superior de Investigaciones Científicas.

Casares, Julio (1958), *Vvedenije v sovremennuju leksikografiju*, Moscú: Izdatel'stvo inostrannoj literatury.

Corpas Pastor, Gloria (1996), *Manual de fraseología española*, Madrid: Gredos.

Corpas Pastor, Gloria (2021), «Constructional idioms of 'insanity' in English and Spanish: A corpus-based study», *Lingua* 254, 1-20.

Corpas Pastor, Gloria – Bautista Zambrana, Mª Rosario – Hidalgo-Ternero, Carlos (eds.) (2021), *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus,* Granada: Comares.

Corpas Pastor, Gloria – Colson, Jean-Pierre (eds.) (2020), *Computational Phraseology* (IVITRA Research in Linguistics and Literature, 24), Ámsterdam – Filadelfia: John Benjamins.

Corpas Pastor, Gloria – Mitkov, Ruslan (eds.) (2019), *Computational and Corpus-Based Phraseology. Third International Conference, Europhras 2019, Malaga, Spain, September 25–27, 2019, Proceedings*, Berlín: Springer.

Corpas Pastor, Gloria – Mitkov, Ruslan (eds.) (2022), *Computational and Corpus-Based Phraseology. Forth International Conference, Europhras 2022, Malaga, Spain, September 28-30, 2022, Proceedings*, Berlín: Springer.

Erman, Britt – Warren, Beatrice (2000), «The idiom principle and the open choice principle», *Text & Talk* 20(1), 29-62.

Gutiérrez Rubio, Enrique (2022), «Frecuencia de uso de locuciones y paremias en el corpus Spanish Web 2018 (esTenTen18): implicaciones didácticas y lexicográficas», en Corpas Pastor *et al.* (eds.), *Proceedings of the International Conference EUROPHRAS 2022, Malaga, Spain, 28-30 September, 2022*, Bulgaria: INCOMA, 26-33.

Mellado Blanco, Carmen – Gutiérrez Rubio, Enrique (2020), «Nuevas aportaciones de la Gramática de Construcciones a los estudios de fraseología en las lenguas románicas», *Romanica Olomucensia* 32/1, 1-12.

Monti, Joanna – Seretan, Violeta – Corpas Pastor, Gloria – Mitkov, Ruslan (eds.) (2018), *Multiword units in machine translation and translation technology* (Current Issues in Linguistic Theory, 341), Ámsterdam – Filadelphia: John Benjamins.

Moon, Rosamund (1998), *Fixed expressions and idioms in English. A corpus-based approach,* Nueva York: Clarendon Press.

Penadés Martínez, Inmaculada (1999), *La enseñanza de las unidades fraseológicas*, Madrid: Arco/Libros.

Penadés Martínez, Inmaculada (2019), *Diccionario de locuciones idiomáticas del español actual* [disponible en <http://www.diccionariodilea.es/diccionario>, 12/6/2023].

Ruiz Gurillo, Leonor (1997), *Aspectos de fraseología teórica española*, Valencia: Universitat de València.

Seco, Manuel – Andrés, Olimpia – Ramos, Gabino (2017), *Diccionario fraseológico documentado del español actual: locuciones y modismos españoles,* Madrid: JdeJ Editores.

Varela, Fernando – Kubarth, Hugo (1994), *Diccionario fraseológico del español moderno*, Madrid: Gredos.

Zuluaga, Alberto (1980), *Introducción al estudio de las expresiones fijas*, Frankfurt am Main: Peter D. Lang.

*Gloria Corpas Pastor (editora invitada)*

*Enrique Gutiérrez Rubio (director)*