[Borrador] Rodríguez-Espiñeira, María José y Hella Olbertz. “Los corpus del español y los estudios de sintaxis funcional”, en Parodi, Giovanni, Pascual Cantos y Chad Howe (eds.): *Lingüística de corpus en español / The Routledge Handbook of Spanish Corpus Linguistics*, cap. 10, pp. 150-162.
DOI: 10.4324/9780429329296-13.
https://www.routledge.com/Linguistica-de-corpus-en-espanol--The-Routledge-Handbook-of-Spanish/Parodi-Cantos-Gomez-Howe/p/book/9780367350123

# Los corpus del español y los estudios de sintaxis funcional

(Spanish corpora and functional studies of syntax)

*María José Rodríguez Espiñeira y Hella Olbertz*

## 1. Introducción

Este capítulo proporciona ejemplos ilustrativos de algunas áreas de la sintaxis del español que están bien representadas en los estudios con corpus. Las técnicas de la lingüística de corpus permiten estudiar la variabilidad e identificar los esquemas de uso más frecuentes. Los trabajos que sirven de ilustración están basados en modelos teóricos de base cognitiva o vinculados con la Gramática de Construcciones. En el capítulo se presenta investigación sintáctica centrada en la valencia cuantitativa y cualitativa de los elementos predicativos, en particular de los verbales. Concretamente, se tienen en cuenta (i) la valencia potencial de los predicados de cognición, comunicación, contacto físico, cambio, percepción, sensación y evaluación; (ii) clases específicas de construcciones, como las transitivas, ditransitivas, transitivas complejas (con predicados secundarios), pasivas, pronominales y construcciones con experimentador. Además de reseñar algunos trabajos basados en muestras obtenidas de concordancias de corpus, se concede atención preferente a estudios originados en bases de datos que proporcionan anotación morfosintáctica, funcional, y/o semántica, como BDS y ADESSE. En cuanto a los métodos usados en los trabajos sintácticos presentados, el núcleo de este capítulo lo constituyen los análisis cualitativos basados en frecuencias simples, si bien se mencionan algunas investigaciones con análisis multifactoriales que miden el peso relativo de factores heterogéneos en las elecciones de los hablantes.

**Palabras clave**: sintaxis basada en corpus, valencia cuantitativa y cualitativa, esquemas semántico-sintácticos, construcciones

## 1. Introduction

This chapter provides illustrative examples of those areas of Spanish syntax in which corpora are widely used. The techniques of corpus linguistics allow for research into variability and for the identification of the patterns most frequently used. The studies that serve to illustrate the application of corpora to this aim are based on cognitive approaches or on Construction Grammar. The chapter presents corpus-based syntactic research with the emphasis on the quantitative and qualitative valency of predicative items, particularly verbal ones. More specifically, it discusses (i) the potential valencies of predicates of cognition, communication,

physical contact, change, perception, sensation and evaluation, and (ii) specific classes of constructions, such as transitives or ditransitives, complex transitive constructions, passives, (pseudo-) reflexives and experiencer constructions. In addition to reviewing a number of studies based on samples taken from corpus concordances, the authors pay particular attention to studies originating in databases that provide morphosyntactic, functional and/or semantic annotation, such as BDS and ADESSE. As regards the methodology used in the syntactic papers presented, qualitative analyses based on simple frequency data at the core of this chapter. However, studies containing multifactorial analyses, which measure the relative weight of heterogeneous factors in the grammatical choices that speakers make, are also taken into account.

**Keywords**: corpus-based syntax, quantitative and qualitative valency, semantic-syntactic patterns, constructions

## 2. Conceptos fundamentales

La investigación sintáctica requiere examinar propiedades semánticas, funcionales y pragmáticas más abstractas que las de tipo formal y distribucional, cada vez más accesibles gracias a los avances en el desarrollo de etiquetadores morfosintácticos. Remontándonos a los inicios de los estudios sintácticos con corpus, los investigadores han optado por: (i) anotar manualmente las propiedades de las unidades que se desea estudiar, como se ha hecho con las bases de datos sintácticas (BDS) y semánticas (ADESSE), comentadas en el apartado 3.1 de este capítulo; (ii) trabajar con anotadores sintácticos automáticos (*treebank*), que permiten recuperar secuencias que satisfacen ciertas condiciones estructurales (*cfr.* Rojo 2016); en ambos casos, el tamaño de los corpus analizados es reducido; (iii) aplicar a un corpus herramientas más avanzadas, que permitan búsquedas combinando diferentes rasgos (*cfr.* §5). El carácter parcial e incompleto de los diccionarios de construcción y régimen existentes para el español explica que un objetivo destacado de la investigación sintáctica haya sido el análisis de la valencia y de las restricciones combinatorias de los argumentos o actantes.

Los estudios de sintaxis funcional parten del presupuesto de que la lengua tiene como función primordial la comunicación interpersonal, por lo que debe ser estudiada y analizada en contextos de uso real. A pesar de que el uso sea inherentemente variable, la estructura lingüística goza de una cierta estabilidad si se la considera desde una perspectiva sincrónica. Las distintas escuelas funcionalistas difieren en el grado de énfasis que otorgan a la variabilidad y a la relativa estabilidad, lo que permite situarlas entre dos polos: el funcionalismo radical, que postula que las lenguas son sistemas dinámicos, que la gramática emerge del uso y responde a presiones discursivas, y los modelos formalizados de gramáticas funcionales, una de cuyas metas es formular una teoría con un elevado grado de adecuación tipológica (Butler 2004, §5). Lo que tienen en común los modelos funcionales es que 'toman las lenguas en serio' (Dik 1997, 17–18), lo que impulsa a trabajar con datos lingüísticos auténticos. La disponibilidad de corpus cada vez más grandes y la mejora de las técnicas de análisis en los últimos treinta años representa un desafío para los modelos teóricos, especialmente para los funcionales, ya que es posible poner a prueba los conceptos teóricos para refinarlos o revisarlos.

En este capítulo, que otorga mayor énfasis a la variabilidad que a la estabilidad lingüística, prestamos atención a los trabajos que examinan las propiedades combinatorias de los verbos con los restantes elementos léxicos y gramaticales, es decir, la valencia predicativa verbal, tanto en su vertiente cuantitativa como cualitativa. La valencia puede servir como punto de partida para el estudio de fenómenos relacionados con la transitividad, la diátesis, la

codificación variable de los argumentos y las diferencias de significado asociadas a subtipos de unidades que alternan en un hueco funcional, como cláusulas finitas, no finitas y reducidas, entre otras.

Adoptando una distinción de Ágel (1995), García-Miguel (2012, 40–45) propone diferenciar entre el potencial valencial de un verbo (o un sentido verbal) y las realizaciones valenciales. Con un ejemplo de este autor, un verbo como *vencer* está vinculado con tres participantes actanciales: Vencedor (A1), Vencido (A2) y Competición (A3). Las realizaciones valenciales, o patrones actanciales, se expresan en diferentes estructuras sintácticas, como las representadas por estos ejemplos: (i) *Los liberales vencieron a la Iglesia;* (ii) *Habían vencido los buenos;* (iii) *Los carlistas vencieron en la batalla de Lácar*. En los ejemplos citados cada esquema sintáctico perfila diferentes facetas de la competición: (i) SUJ (Vencedor) –V (*vencer*) – OD (Vencido); (ii) V(*vencer*) – SUJ (Vencedor); (iii) SUJ (Vencedor) –V (*vencer*) – CPREP (Competición).[1] Los datos de corpus permiten constatar que es infrecuente que un verbo se asocie con una única estructura sintáctica y con los mismos participantes.

Cuando la valencia de un predicado se concibe como una generalización a partir del uso registrado, en el potencial valencial de los predicados tienen cabida diversos tipos de actantes (García-Miguel 2007; González Domínguez 2014): a) Los inherentes o esenciales vinculados con el proceso designado, como el móvil con los verbos de desplazamiento, o el poseedor y lo poseído con verbos de posesión; b) Los implícitos o recuperables del contexto; c) Los evocados dentro de una construcción específica, como los agentes en las construcciones medio-pasivas; d) Los opcionales o débilmente evocados, como el origen con los verbos de desplazamiento; e) Los adicionales, como sucede cuando se hacen prominentes elementos concretos de las cadenas causales (*v.gr.* el instrumento) o se desdoblan facetas de un argumento (*v.gr.* dativos posesivos).

Los enfoques basados en el uso tratan de conciliar dos puntos de vista opuestos sobre la asociación entre verbos y construcciones: a) a través de los verbos se pueden identificar alternancias de esquemas sintáctico-semánticos; b) gracias a las construcciones se pueden formular generalizaciones sobre los esquemas sintáctico-semánticos con diferentes verbos. A medida que se han ido sucediendo diferentes versiones de la Gramática de Construcciones, se ha generalizado la noción de *construcción* para diferentes tipos de unidades simbólicas: el concepto se aplica a cualquier patrón de asociación entre expresión y contenido. Con un ejemplo adaptado de Croft (2009, 477), la frase idiomática [SUJ *estirar la pata*] es una unidad sintáctica mínima e idiosincrásica. La construcción más esquemática específica de un verbo, como [SUJ *estirar* OD], es un esquema argumental. La construcción completamente esquemática [SUJ Vtr OD] representa la cláusula transitiva, una realización de una construcción argumental, que puede describirse exclusivamente mediante su estructura sintáctica. La construcción argumental es un tipo específico de construcción, que proporciona los mecanismos básicos de expresión para las cláusulas de una lengua. Como vimos en los párrafos anteriores, la forma de esta construcción depende de varios factores sintácticos, semánticos y contextuales.

Por otro lado, la lingüística cognitiva se ha revelado fructífera para explicar los esquemas sintácticos alternativos de un verbo, sin cambios drásticos de acepción. En su artículo sobre el 'punto de referencia' Langacker (1993) argumenta que existe una tendencia a seleccionar una entidad cognitivamente prominente con el fin de evocar otra entidad relacionada menos sobresaliente. Son ilustrativas al respecto las relaciones posesivas, en las que el poseedor humano actúa como punto de referencia para establecer contacto mental con el objeto de posesión (*Le tapó la cara*). También es posible que el acceso mental a un evento se efectúe mediante la topicalización de un individuo que participa en la escena verbal, como sucede en las construcciones transitivas complejas del tipo *Jano lo vio más sereno* o *Papá me*

*cree muerta* (SUJ-OD-predicativo del OD). En este tipo de construcción, el fenómeno percibido o el contenido conceptualizado se disocia en dos constituyentes: un objeto directo y un complemento predicativo del objeto. La elección entre este esquema y la cláusula finita (*Jano vio que estaba más sereno; Papá creía que estaba muerta*) no es aleatoria, ya que la construcción transitiva compleja está sujeta a muchas restricciones (García-Miguel y Comesaña 2004, 409; Gonzálvez-García 2003, 42–43).[2]

En resumen, en los marcos funcionales se defiende que las alternativas construccionales no guardan entre sí relaciones puramente formales, por lo que se intenta explorar qué tipo de motivaciones semánticas o discursivas corresponden a cada esquema constructivo. En este sentido, se postula que la gramática es un mecanismo que codifica tanto información semántica como pragmático-discursiva. Esto explica por qué varios trabajos reseñados en este capítulo aluden a la relación entre la codificación sintáctica y las propiedades semánticas y pragmático-discursivas. Es este un dominio en el que la lingüística de corpus puede aportar conocimiento fundado, dado que algunas expresiones juzgadas como agramaticales o dudosas no lo son provistas del contexto adecuado (Fernández 2007, 13, 48; Vaamonde 2011, 292).

## 3. Estado de la cuestión

En este apartado mostraremos algunas contribuciones basadas en datos obtenidos de corpus específicos: ARTHUS y BDS; ADESSE; CREA y otros recursos.

### *3.1. ARTHUS y BDS*

A comienzos de los años 90 se creó el corpus ARTHUS (Archivo de Textos Hispánicos de la Universidad de Santiago), integrado por textos mayoritariamente escritos (narrativa, teatro, ensayo, prensa) publicados entre 1981 y 1991, compuesto de 1,5 millones de palabras. Esta iniciativa fue una de entre otras varias desarrolladas en el período que precede a la publicación de los grandes corpus sincrónicos (CREA, CORPES) y diacrónicos (CORDE) de la Real Academia Española, por un lado, y el Corpus del Español de Mark Davies, por otro (Rojo 2016).

Lo que distingue el corpus ARTHUS de los demás es su elaboración posterior, que brinda al usuario una visión detallada y exhaustiva de la valencia sintáctica y semántica de los predicados verbales en el español actual. En una primera etapa, se elabora en detalle la estructura sintáctica. Tras el análisis manual de 158 624 cláusulas, los resultados fueron almacenados en la BDS (Base de Datos Sintácticos del español), compuesta por 63 campos. En la BDS, el texto y la anotación sintáctica no están integrados, sino que se presentan en paralelo, uno de los rasgos que distinguen a este recurso de un corpus anotado sintácticamente. Este análisis presupone una agrupación de elementos en clases funcionales preestablecidas (p. ej. sujeto, complementos directo e indirecto, predicativos), pero este posible inconveniente se compensa con el amplio abanico de rasgos semánticos y sintácticos anotados para cada hueco funcional: categoría sintáctica (subtipos de palabras, de frases y de cláusulas), tipo de referente (±animado, ±concreto), preposición, definitud, persona y número de los clíticos, entre otros. En un nivel más refinado, cada cláusula se caracteriza según parámetros específicos, p. ej. voz del predicado, polaridad, modalidad, orden de constituyentes; para las cláusulas subordinadas, propiedades integrativas, tipo de nexo introductor y función desempeñada. La aplicación de consulta (www.bds.usc.es/) permite acceder a los (sub)esquemas sintácticos en que aparece un verbo, a los verbos que se registran

en un (sub)esquema sintáctico y a los ejemplos de un verbo en un (sub)esquema específico (Rojo 2001).

Los datos de la BDS tuvieron una explotación solo parcial en algunos trabajos publicados y se usaron para extraer ejemplos documentados, obtener recuentos sobre la frecuencia de esquemas sintácticos (*cfr.* Rojo 2011) y determinar las condiciones de uso y el rendimiento comunicativo de diferentes opciones semánticas, sintácticas y pragmáticas. Entre las investigaciones sintácticas basadas en datos proporcionados por este recurso, reseñamos tres trabajos sobre (i) orden de constituyentes; (ii) transitividad y participantes centrales; (iii) construcciones pasivas.

(i) López Meirama (1997) revisa la 'hipótesis ergativa', según la cual los verbos intransitivos inacusativos, como *abundar, aparecer, llegar, ocurrir, salir* o *surgir*, imponen al sujeto la posición posverbal. El estudio constata que la mayoría de los verbos son polisémicos; por ejemplo, los de movimiento direccional admiten tanto sujetos animados (agentivos e inagentivos) como inanimados: *El hijo sale y vuelve pronto / Abrí el grifo y salió el agua tibia; El reyezuelo de la casa asoma dando gritos / Asoma la sonrisa horrible; Tía Delia vino enseguida / Entonces empezaron a venir los hijos*. Con respecto al orden de constituyentes, es más apropiado hablar de construcciones inacusativas e inergativas que hacerlo de verbos. Además, los datos indican que la posposición es un fenómeno complejo, en el que junto al carácter inagentivo del sujeto intervienen otros factores, como la animación o la definitud: alrededor del 85% de los sujetos preverbales de las cláusulas intransitivas del corpus ocupan posiciones altas en la jerarquía de referencialidad y definitud, frente al 90% de los sujetos posverbales, con posiciones bajas en dicha escala.

(ii) El trabajo de Vázquez Rozas (2006) detecta algunos problemas en la hipótesis de transitividad de Hopper y Thompson (1980) a partir de un análisis que contrasta los verbos psicológicos con experimentador en dativo (del tipo *gustar*) con los verbos transitivos correspondientes (del tipo *amar*). La autora proporciona argumentos formales y semánticos para considerar al objeto indirecto (dativo) del español como una función central de la cláusula y demuestra que las cláusulas con SUJ-V-OI se desvían del prototipo transitivo (con sujeto agentivo y animado en un 85,76% de cláusulas en la BDS). En contraste con las cláusulas transitivas, las del esquema SUJ-V-OI ofrecen la tendencia inversa en cuanto a la animación del sujeto: un 71,65% son inanimados, frente al 28,35% de animados.

(iii) El estudio de Fernández (2007) consiste en un análisis discursivo de la voz pasiva en español, en el marco de un modelo cognitivo-funcional. La autora lleva a cabo cálculos de distancia referencial y de persistencia de tópico. Los datos muestran que la función discursiva principal de la pasiva perifrástica (92% de los casos) consiste en mantener como sujeto al participante con mayor continuidad discursiva en el contexto. Por ello, los sujetos de la pasiva perifrástica son, en su mayor parte, definidos, con referente específico. En un 8% de casos, este tipo de pasiva tiene como función textual secundaria la "simple crónica de sucesos", que es la principal función discursiva de la pasiva con *se*. El hecho de que se produzca una "(parcial) convergencia funcional" favorece que ambas construcciones sean intercambiables para muchos hablantes de español. Las pasivas reflejas prefieren el orden VS, aparecen en enunciados téticos que expresan la ocurrencia de un suceso y muestran predilección por participantes más inespecíficos. Cuando tienen lectura genérica no formulan generalizaciones fruto de la experiencia directa del hablante, sino que incluyen voces de terceros, en tanto que aluden a normas sociales, métodos establecidos, rutinas, o simplemente información "de segunda mano". Por otra parte, las construcciones impersonales con *uno, tú/vos* presentan una generalización a partir de una experiencia personal. En cuanto a la construcción impersonal con *se*, presenta un evento con un bajo nivel de elaboración y guarda relación con la idea de habitualidad.

### *3.2. ADESSE*

A partir de 2002 se desarrolla en la Universidad de Vigo el proyecto ADESSE (Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español), que consiste en el enriquecimiento de los datos de la BDS con la semántica léxica correspondiente a 3 344 verbos, con diferenciación de acepciones o sentidos y el añadido de los papeles semánticos de los argumentos. Los verbos se categorizan en seis grupos semánticos mayores según la tipología de procesos de Halliday (1985), con subcategorizaciones adicionales cada vez más refinadas (p. ej. la categoría Material contiene la subcategoría Espacio, que, a su vez se subdivide en, entre otras, Movimiento, Lugar, Orientación). En la base de datos ADESSE (www.adesse.uvigo.es) se distingue entre el potencial valencial y las realizaciones valenciales y se proporciona el perfil combinatorio de los esquemas sintácticos. Para más detalles sobre ADESSE, *cfr*. Vaamonde *et al.* (2010) y García-Miguel (2012, 2014).

Los trabajos basados en este recurso están dedicados, fundamentalmente, al examen de la polisemia verbal, de los patrones valenciales y sus realizaciones morfosintácticas (García-Miguel 2007), así como al análisis de grupos específicos de predicados, como los de cognición, (García-Miguel y Comesaña 2004), competición (García-Miguel 2014) o contacto físico (González Domínguez 2014). En todas estas obras se siguen los principios mencionados en el §2, ilustrados con alguno de estos trabajos; nos detendremos brevemente en dos de ellos. En González Domínguez (2014) se establecen similitudes y diferencias entre varias subclases de verbos de contacto identificados en ADESSE, tomando en cuenta rasgos como el papel del movimiento (±dinámico), el de la fuerza (±impacto) y la transmisión de energía (±simétrica) en los usos recíprocos. Los verbos de contacto sin impacto se subdividen en cuatro subclases: maneras de tocar, contacto táctil, con fricción y afectivo. Los de contacto con impacto se distribuyen en dos subclases: plenos y ligeros. También se proporcionan las relaciones semánticas entre los usos literales y figurados, en forma de redes conceptuales basadas en procesos de especificación, generalización, metáfora y metonimia.

La tesis de Vaamonde (2011) se ocupa de tres esquemas sintácticos, SUJ-V-OD, SUJ-V-OD-OI, SUJ-V-OD-CPREP, uno de cuyos huecos funcionales es ocupado por un nombre de parte del cuerpo. Para llevar a cabo la investigación fue preciso identificar y anotar manualmente todos los argumentos que designan una parte del cuerpo o un poseedor, junto a otros aspectos relevantes, como la presencia de modificadores descriptivos o la de adverbios de manera (Vaamonde 2011, 102–103). Además de confirmar que la construcción de dativo posesivo (*Le besó la mejilla*) es la opción más compleja, más frecuente y menos marcada (510 ejemplos con clítico), el trabajo demuestra que la construcción con posesivo interno (*Besó* {*su mejilla / la mejilla de Ana*}) es más simple y marcada, y se documenta exclusivamente en la lengua literaria (361 ejemplos con posesivo o modificador adnominal). Haciendo uso del concepto de 'punto de referencia' de Langacker (1993) mencionado en §2 arriba, Vaamonde (2011, §5.5) explica con detalle las razones estilísticas que la justifican, relativas a la prominencia de la entidad poseída en un contexto determinado. La opción más frecuente puede "resultar demasiado convencional", por lo que el uso del posesivo interno obedece al intento de "romper el patrón establecido y dotar a la acción de una mayor fuerza expresiva" (2011, 389).

### *3.3. CREA y otros recursos*

La aparición del *Corpus del español* de Mark Davies, primero, y la creación de los corpus académicos, después, contribuyen decisivamente a reorientar los estudios de sintaxis al

proporcionar fácil acceso a datos lingüísticos obtenidos en contextos naturales de uso. Desde entonces es posible poner el foco en lo que es frecuente e infrecuente, típico o atípico, y dirigir la mirada hacia los factores que condicionan o explican las diferentes realizaciones de los esquemas sintácticos. La repercusión de este enfoque es perceptible también en los modelos teóricos no funcionales, que han pasado de trabajar casi exclusivamente con datos introspectivos a tener en cuenta la variación sintáctica observada en textos producidos por diferentes emisores.[3]

En este subapartado centramos nuestra atención en trabajos elaborados fuera de España. Primero consideramos las actividades desarrolladas en las universidades de Lovaina y Gante, para después presentar un trabajo publicado en México. En Lovaina fue pionero en el estudio con corpus el hispanista Josse de Kock y su trabajo tuvo continuidad en los estudios de orientación cognitiva llevados a cabo por Nicole Delbecque y sus discípulos (con fuerte impronta semántica). En Gante, fue Eugeen Roegiest el impulsor de los estudios gramaticales con corpus, a menudo con enfoque contrastivo de dos o más lenguas romances. Varios verbos de percepción han acaparado el interés de ambos grupos.

Enghels y Roegiest (2004) examinan las diferencias físicas y cognitivas entre la percepción visual (*ver*) y la auditiva (*oír*), particularmente cuando se combinan con complementos en infinitivo: dimensión espacial/temporal, movilidad del órgano perceptor, distancia entre perceptor y percepto. Y comprueban que esas diferencias pueden explicar el mayor uso de la preposición *a* y del clítico *le* con el verbo *oír*. En el dominio sensorial, un perceptor visual es fuente primaria de información, en tanto que testigo directo (*Lo vi entrar, Le vi volver el rostro*), pero el que oye puede ser fuente primaria (*Lo oí llorar*) o secundaria (*Eso le oí decir a la patrona*).[4]

Hanegreefs (2006) ofrece un análisis empírico del verbo *mirar*, agentivo de percepción física visual, con datos del Corpus del Español y del CREA. La construcción más frecuente de este verbo es la transitiva (60% en el corpus analizado), pero también tiene peso la construcción intransitiva con un sintagma preposicional locativo, lo que hace de *mirar* un verbo afín a los de movimiento. El segundo argumento de *mirar* admite dos formatos alternativos: sintagma nominal (*Se detiene y mira atentamente <u>el suelo</u>*; *Francesillo mira las cartas y mira <u>los ojos de Víctor</u>*) o sintagma preposicional (*Caminaba y miraba <u>al suelo</u>*; *Él la mira directamente <u>a los ojos</u>*). En el primer caso el objetivo de la mirada se asimila a un objeto perceptible (OD), mientras que en el segundo se perfila la orientación espacial (CPREP).

La investigación de Hanegreefs (2008) es un exhaustivo estudio de 18 verbos de percepción, basado en el uso registrado en español de España (CREA). La autora pone especial énfasis en explorar las diferencias conceptuales y semánticas entre los verbos y en determinar de qué modo la selección del tipo de complemento está condicionada por la forma en que el conceptualizador organiza o construye la escena de percepción.

Por lo que respecta a la producción americana, una parte considerable del trabajo con corpus tiene orientación diacrónica, sociolingüística, o discursivo-textual (*cfr*. Travis y Torres Cacoullos 2012), por lo que en este capítulo reseñaremos un solo estudio realizado en la UNAM. En Melis (2011) se analizan, con datos del CREA, tres verbos de suficiencia (*bastar*, *faltar* y *sobrar*), que comparten la expresión de un juicio valorativo y cuya valencia se discute en la bibliografía: monovalentes o bivalentes. En el dominio de la ‘suficiencia’, las entidades codificadas como sujeto son evaluadas con respecto a una magnitud determinada, que sirve como punto de anclaje para el juicio modal –la cantidad que el hablante considera necesaria– (*No bastan la tiza y los libros para ejercer la docencia*). Los verbos *faltar* y *sobrar* se emplean para indicar que la magnitud expresada por una entidad no armoniza con la cantidad considerada necesaria, por insuficiente o por excesiva. Los datos de corpus examinados sugieren que debe considerarse valencial, además de la entidad evaluada, el

constituyente que adopta la forma de un complemento final (*Este dinero basta para comprar la comida*), considerado como "criterio pragmático". Melis sostiene que este elemento designa la situación con respecto a la cual cobra sentido el juicio de suficiencia implicado en el significado verbal. En los enunciados donde coaparecen tanto un dativo como un complemento con *para* (*Contaba los días que me faltaban para hacer yo mismo ese viaje*), el criterio pragmático se desdobla. El dativo funciona como 'punto de referencia' (Langacker 1993), por tratarse de un participante individualizado y prominente; el proceso de topicalización (elevación) permite que el hablante formule su juicio de suficiencia desde la perspectiva de la entidad (humana) a la que le concierne dicho juicio. Esta relación tan estrecha entre la entidad evaluada y el referente en dativo permite no incluir el complemento con *para* (*No le faltaban razones*), que puede inferirse o recuperarse contextualmente. Cuando la entidad evaluada y el dativo entablan una relación (posesión, propiedad o cualidad humana), la construcción con los verbos *sobrar* y *faltar* se desliza desde el campo de la existencia hacia el de atribución de propiedades (*Tienes alguna simpatía. Pero te faltan lecturas*), lo que convierte al dativo en imprescindible.

## 4. Consideraciones metodológicas

Como indica Rojo (2014), el perfil ideal de una investigación sintáctica basada en corpus reúne las siguientes características: a) los datos se obtienen de contextos naturales de uso y se toman como representativos de una lengua o de una variedad de lengua; b) los datos no son sometidos a filtros o selecciones previas, sino que se intenta llevar a cabo un análisis exhaustivo del corpus completo (principio de *total accountability*) o de un subconjunto obtenido de forma aleatoria; c) se usan métodos cuantitativos (frecuencias, probabilidades) combinados con análisis cualitativos; d) la descripción tiene en cuenta patrones de coaparición, por lo que el contexto resulta decisivo. Una exigencia derivada del uso de corpus textuales es la exhaustividad, que implica "una revisión continua de las categorías cualitativas preestablecidas y un movimiento continuo de ida y vuelta de los datos a la teoría y viceversa" (García-Miguel 2012, 32). Las investigaciones sintácticas deben someter los datos analizados a parametrización sintáctica, semántica e incluso discursiva, lo que impide satisfacer todos los requisitos mencionados. Por este motivo, muy a menudo: (i) se trabaja con repertorios de textos creados con fines específicos para describir una variedad dialectal, un tipo de texto o discurso, la lengua conversacional, etc. (ii) se confeccionan muestras o subcorpus más pequeños, que resultan más manejables; así, para obtener el perfil combinatorio del verbo *sentir*, Jansegers *et al.* (2015) manejan el corpus CREA, al que aplican inicialmente dos filtros: temporal (2000–2004) y geográfico (España); para facilitar la clasificación manual de los datos, que tiene en cuenta 32 rasgos, seleccionan una muestra aleatoria del 25% de las concordancias, que completan con datos orales de otros corpus.

Desde el libro de Tognini-Bonelli (2001) es habitual contraponer dos puntos de vista en la lingüística de corpus: a) el enfoque basado en corpus (*corpus-based approach*) y b) el enfoque guiado por el corpus (*corpus-driven approach*). El primero se caracteriza por el empleo de datos auténticos en vez de inventados para apoyar una teoría e implica partir de categorías descriptivas preestablecidas; el segundo parte de los datos de corpus y trata de obtener las categorías descriptivas de los datos. Muchos trabajos de sintaxis con corpus adoptan la primera perspectiva: el corpus se emplea para refinar la descripción o añadir una dimensión cuantitativa, por lo que a veces los datos incómodos son relegados o se les asigna un papel secundario en la descripción (Tognini-Bonelli 2001, 72–74; Butler 2004, §4). García-Miguel (2012, 32–33) sostiene que puede resultar artificial la contraposición entre estos dos enfoques, porque hasta hace muy poco se ha trabajado mayoritariamente con corpus

no analizados y no anotados, lo cual limita las búsquedas a palabras ortográficas y permite únicamente estudios sobre ejemplos contextualizados o sobre combinaciones frecuentes de palabras ("colocaciones"). Pero existe otro factor importante, ya mencionado en §2, vinculado con el tipo de investigación: la sintaxis requiere estudiar propiedades más abstractas que las que ofrece un texto plano, por lo que ni siquiera son suficientes las anotaciones que proporciona la etiquetación morfosintáctica.[5]

De acuerdo con el tipo de datos y la complejidad estadística de las mediciones llevadas a cabo, los trabajos de sintaxis se pueden agrupar en tres categorías (Gries 2013): (i) los que aportan frecuencias absolutas y probabilidades de aparición de los (elementos de) los esquemas sintácticos; (ii) los que miden la fuerza de la asociación entre (clases de) unidades léxicas y construcciones y cuantifican su grado de atracción o de repulsión, según las pautas del "análisis colostruccional"; [6] (iii) los que tratan de evaluar la importancia relativa de distintos factores lingüísticos mediante análisis multifactoriales, que son particularmente necesarios cuando los valores obtenidos previamente para cada factor parecen tener un peso similar. En adelante ilustramos cada una de estas categorías con ejemplos.

(i) Los recuentos de frecuencias no bastan para explicar la combinación de unidades, pero permiten identificar patrones de uso y fenómenos no tenidos en cuenta que merecen ser interpretados y justificados. Como ejemplo sirvan los verbos de actitud proposicional (*creer, notar, pensar, recordar* o *saber*). El hecho de que tengan como construcción básica y más frecuente la monotransitiva ha llevado a postular que está bloqueada la presencia de un complemento indirecto. Sin embargo, García-Miguel y Comesaña (2004) comprueban que 25 verbos de la clase se registran también en el esquema ditransitivo, aunque con una frecuencia considerablemente menor. Entre las motivaciones que justifican la presencia de un objeto indirecto con estos verbos, García-Miguel y Comesaña destacan dos: a) las extensiones metafóricas de transferencia y comunicación, por lo que los verbos cognitivos adoptan en el esquema ditransitivo significados causativos (*Todo me evocaba la imagen de Beatriz; Le recordaba a su madre*, 'hacer que recuerde') o de comunicación (*Le recordé que era lunes* 'decirle a alguien'); b) la expresión de la adquisición por el sujeto de un conocimiento o una creencia cuya fuente es el objeto indirecto (*No le creí una sola palabra*).

El estudio de Vázquez Rozas y Miglio (2006), sobre la construcción biactancial con Experimentador y Estímulo en español e italiano, arroja luz sobre las frecuencias reales de los esquemas transitivo (*Yo detestaba a los hombres altaneros*) e intransitivo (*Le gustaban las fiestas ruidosas y largas*) y sobre su función discursiva. En el trabajo se aplican a ambas construcciones pruebas estadísticas (árboles de clasificación y regresión) con diferentes predictores: animación del Estímulo, persona y número del Experimentador y género textual.

(ii) Rojo (2011) emplea el "análisis de colexemas" para examinar la predilección o repugnancia de 25 verbos por el esquema sintáctico ditransitivo SUJ-V-OD-OI. Fuera de cuestiones de detalle, observa que los rangos derivados de la prueba del $\chi^2$ y del índice de Fisher resultan bastante congruentes, frente a la disparidad obtenida al ordenar los verbos con las frecuencias (total y normalizada) de cada verbo en el esquema y en el total de usos en la BDS.

Pedersen (2016) analiza la coaparición de la construcción esquemática 'evento de movimiento télico' –[SUJ-V-*a* SN]– con 249 verbos de movimiento y ordena los lexemas según su atracción por la construcción: a) los prototípicos del esquema (los de trayectoria con punto final, como *llegar*); b) los que muestran una asociación débil (algunos de manera de movimiento, como *correr*) y verbos que no están asociados con la construcción (de trayectoria, como *embarcarse* o *atracar*, o de manera de movimiento, como *cojear*). El trabajo ofrece información cualitativa sobre la variación detectada, al distinguir entre variación descartable (escasos ejemplos documentados) y variación disponible (más ejemplos documentados).

Yoon y Wulff (2016) examinan la alternancia entre complementos con infinitivo –561 casos– y con cláusula flexionada (*que*) –795 casos– en un corpus de prosa periodística (AnCoraEs, véase Rojo 2016). Mediante un análisis distintivo de colexemas, los autores identifican los verbos dominantes vinculados con cada tipo de complemento: los infinitivos son atraídos por verbos de deseo (p. ej. *querer*, *intentar*, *decidir*, *pretender*, *preferir*) y las cláusulas flexionadas se asocian con verbos de comunicación (*decir*, *explicar*, *anunciar*, entre otros) y de actividad mental (p. ej. *creer*, *recordar*, *reconocer*, *entender*). El mismo tipo de análisis aplicado a los verbos dependientes no ofrece resultados tan concluyentes: los lexemas en cláusulas de infinitivo forman una clase heterogénea, sin vínculos específicos con la construcción. En cambio, los lexemas verbales presentes en los complementos flexionados son más homogéneos, pues los más frecuentes son verbos livianos del tipo *ser, haber, estar, ir, poder*.

Según Granvik (2015, §3.1) el *análisis colostruccional* presenta algunas limitaciones cuando se intenta aplicar a la construcción encapsuladora *N* {*que/de que/Ø*} *cláusula completiva* ("Tenemos el *convencimiento (de) que no queda otra salida / de haber elegido la solución más adecuada*"). Por una parte, es difícil identificar en corpus el elenco de sustantivos posibles y, por otra, los valores estadísticos obtenidos son tan reducidos que no permiten clasificar los sustantivos. Para dar cuenta de la atracción o repulsión existente entre un N y la construcción encapsuladora, Granvik (2015) opta por medir tres valores específicos: a) *Dependencia* (el grado en que un N concreto depende de la construcción N + *de que*), que resulta de dividir el número de ejemplos del sustantivo en la construcción por el número total de casos en el corpus; b) *Atracción*, que se obtiene al dividir el número total de usos del sustantivo en la construcción por el número total de usos de la construcción; c) *Razón de momios* (*odds ratio*), que se define como "la posibilidad de que un lexema se presente en una construcción frente al riesgo de que ocurra en otra" (Granvik 2015, 361).

(iii) Para comparar el uso del sujeto léxico con infinitivo en español y portugués, Vanderschueren (2013, §5.27) aplica un análisis de regresión logística binaria que mide la elección entre infinitivo flexionado y no flexionado en portugués teniendo en cuenta varios factores, como el aspecto léxico de la cláusula, el tipo de conector que introduce la cláusula, la presencia de negación, la existencia o no de perífrasis, la (in)existencia de pausa, la distancia entre el infinitivo y el antecedente explícito más cercano al que remite, así como el número de palabras de la cláusula de infinitivo. El análisis multifactorial induce a descartar como factor de selección la presencia de otra cláusula dentro de la de infinitivo. En contextos de correferencia entre el sujeto del infinitivo y el de la cláusula dominante, los factores más relevantes son el grado de autonomía de la cláusula de infinitivo y su grado de verbalidad (presencia de perífrasis, negación, clíticos, formas pronominales, etc.).

## 5. Direcciones futuras y conclusiones

Es innegable que la investigación sintáctica ha ganado en adecuación descriptiva con el apoyo de datos de corpus, como se desprende de los resultados de los trabajos reseñados en este capítulo y de muchos otros que, por razones de espacio, no se han podido mencionar. Además de estar sustentadas en datos de frecuencia de uso, las investigaciones seleccionadas ofrecen explicaciones de tipo cognitivo y comunicativo (funciones discursivas o estructura informativa) y tienen en cuenta diferencias entre la lengua oral y la escrita, así como entre géneros textuales.

Para llevar a cabo estudios cuantitativos, es preciso operacionalizar los criterios de análisis, lo que implica tomar decisiones subjetivas. Una ventaja de los trabajos empíricos es que pueden repetirse, bien con criterios nuevos y el mismo grupo de datos, si se quiere

confirmar la precisión del análisis, bien con otro subconjunto de datos, para medir la representatividad de la muestra. Para ello se requiere que los estudios ofrezcan indicaciones detalladas del procedimiento de selección de datos, de las variables analizadas y del método estadístico empleado, un denominador común en varios estudios sintácticos funcionales reseñados.

La investigación sintáctica sobre (grandes) datos de corpus se enfrenta a muchos retos, pues todavía consume mucho trabajo y tiempo el procesamiento cualitativo de los resultados en bruto, ya que es preciso expurgar las coincidencias recuperadas tras una búsqueda específica. Sería deseable que en el futuro los corpus, o las herramientas creadas para manipularlos, proporcionasen un mayor refinamiento en los metadatos, mejoras en la etiquetación morfosintáctica e incluso una codificación más sutil que incluyese anotaciones semánticas, pragmáticas y prosódicas (para la lengua oral). La herramienta multilingüe Sketch Engine (https://www.sketchengine.eu/) representa un avance evidente, ya que dispone de módulos de etiquetado automático y ofrece varias funciones para el análisis lingüístico, pero el acceso no es libre, al tratarse de un proyecto comercial. Por otra parte, el uso de métodos analíticos multifactoriales compagina bien con la multidimensionalidad de los fenómenos lingüísticos, por lo cual las técnicas cuantitativas y los correspondientes programas de estadística –tales como *R* (https://www.r-project.org/), creado sin fin lucrativo– deberían incluirse en los currículos formativos de los especialistas en lingüística.

Es previsible que en el futuro se incrementen los estudios basados en datos de lengua oral, en particular de la lengua conversacional. Además de confeccionarse corpus más extensos de esta modalidad discursiva, convendría mejorar las técnicas de observación de la lengua hablada, hasta hace muy poco centradas principalmente en la entrevista sociolingüística, con todos sus inconvenientes (*cfr*. Recalde y Vázquez Rozas 2009). El sesgo escrito de muchos corpus o la combinación de diversos géneros discursivos bajo el paraguas de lo oral suponen claras desventajas si lo que se persigue es observar el uso espontáneo de la lengua, un postulado central de los modelos basados en el uso (*cfr*. Travis y Torres Cacoullos 2012) o determinar cómo influye el contexto situacional en la elección de ciertos rasgos gramaticales.

El empleo de corpus no resuelve todos los problemas de carácter teórico o descriptivo que surgen en el curso de una investigación sintáctica, pero proporciona un punto de partida sólido para los análisis y permite someter a prueba la validez de los principios teóricos y descriptivos que los sustentan. Por otra parte, el trabajo con corpus no identifica combinaciones imposibles, pero minimiza el riesgo de atribuir esa interpretación a secuencias que usan los hablantes en diferentes contextos.

## Notas

[1] Las abreviaturas utilizadas son: CPREP (complemento preposicional), OD (objeto directo), OI (objeto indirecto), SN (sintagma nominal), SUJ (sujeto), tr (transitivo), V (verbo).

[2] En Rodríguez Espiñeira (2006) se defiende que la relación entre el OD (base de predicación) y el predicativo no constituye un contenido proposicional.

[3] Butler (2004, 150) opone el discutible papel de la introspección como como fuente única o primaria de datos y su innegable valor para formular hipótesis y analizar los datos, *cfr*. también Fernández (2007, §5.1), Vanderschueren (2013, 7–8).

[4] Véase Rodríguez Espiñeira (2000) para las diferencias entre percepción directa e indirecta en español.

[5] Tognini-Bonelli (2001: 72–74) señala que la etiquetación morfosintáctica, aunque facilita considerablemente el trabajo del lingüista, también "impone un punto de vista específico sobre los datos" (Butler 2004: 153–154).

[6] El análisis estadístico está basado en el Test Exacto de Fisher: mediante una tabla de contingencias (2x2), se especifican las frecuencias por separado y combinadas de un lexema determinado y de la construcción en la que participa. A partir de estas cifras calcula la probabilidad de que la distribución dependa o no del azar. La

repetición de este cálculo con todos los lemas que pueden aparecer en la construcción proporciona una lista de valores *p* ordenados jerárquicamente según la fuerza de asociación que muestren con la construcción.

## Bibliografía citada


Ágel, V. 1995. "Valenzrealisierung, Grammatik und Valenz." *Zeitschrift für germanistische Linguistik* 23: 2–32.

Butler, C. 2004. "Corpus Studies and Functional Linguistic Theories." *Functions of Language* 11 (2): 147–186.

Croft, W. 2009. "Construction Grammar." *Handbook of Cognitive Linguistics*, eds. D. Geeraerts y H. Cuyckens, 463–508. Oxford: Oxford University Press.

Dik, S.C. 1997. *The Theory of Functional Grammar.* Vo1. 1. *The Structure of the Clause*, ed. K. Hengeveld, Berlín: Mouton de Gruyter.

Enghels, R. y E. Roegiest. 2004. "Percepción visual y percepción auditiva: la naturaleza del objeto." En *Cognición y percepción lingüísticas*, eds. E. Serra y G. Wotjak, 47–59, València/Leipzig: Universitat de València/Universität Leipzig.

Fernández, S.S. 2007. *La voz pasiva en español: un análisis discursivo*. Frankfurt am Main: Peter Lang.

García-Miguel, J.M. 2007. "Potencial valencial y tipología de argumentos." En *Perspectivas de análisis de la unidad verbal*, eds. I. Castellón y A. Fernández, 21–33, Barcelona: Universitat de Barcelona.

García-Miguel, J.M. 2012. "Lingüística de corpus y valencia verbal." En *Encoding the Past, Decoding the Future: Corpora in the 21st Century*, eds. I. Moskowich y B. Crespo, 29–57. Cambridge: Cambridge Scholars.

García-Miguel, J.M. 2014. "El perfil combinatorio de los verbos en ADESSE. Polisemia y parasinonimia de verbos de competición." En *Léxico, didáctica y nuevas tecnologías*, ed. Y. Morimoto, 11–37. A Coruña: Universidade da Coruña. Anexos *Revista de Lexicografía* 29.

García-Miguel, J.M. y Comesaña, S. (2004): "Verbs of cognition in Spanish: Constructional schemas and reference points." En *Linguagem, Cultura e Cognição: Estudos de Linguística Cognitiva*, eds. A. Silva, A. Torres y M. Gonçalves, vol. 1, 399–420. Coimbra: Almedina.

González Domínguez, J. 2014. *Análisis léxico y construccional de los verbos de contacto en español*. Tesis doctoral. Universidade de Vigo.

Gonzálvez-García, F. 2003. "Reconstructing object complements in English and Spanish." En *Gramática de Construcciones: contrastes entre el inglés y el español*, ed. M. Martínez Vázquez, 17–58. Huelva: Grupo de Investigación Gramática Contrastiva.

Granvik, A. 2015. "Oraciones completivas de sustantivo: un análisis contrastivo entre portugués y español." *Verba* 42: 347–401.

Gries, S.Th. 2013. "Data in construction grammar." En *The Oxford Handbook of Construction grammar*, ed. G. Trousdale y Th. Hoffmann, 93–108. Oxford: Oxford University Press.

Halliday, M.A.K. 1985. *An Introduction to Functional Grammar.* London: Edward Arnold.

Hanegreefs, H. 2006. "La construcción preposicional con *mirar*: análisis semántico-sintáctico." *Boletín de Lingüística* 18 (25): 22–65.

Hanegreefs, H. 2008. *Los verbos de percepción visual: un análisis de corpus en un marco cognitivo*. Tesis doctoral. Universidad Católica de Lovaina.

Hopper, P.J. y S.A. Thompson. 1980. "Transitivity in Grammar and Discourse." *Language* 56: 251–299.

Jansegers, M., C. Vanderschueren, y R. Enghels. 2015. “The Polysemy of the Spanish Verb *sentir*: A Behavioral Profile Analysis.” *Cognitive Linguistics* 26 (3): 381–421.
Langacker, Ronald. 1993. “Reference Point Constructions.” *Cognitive Linguistics* 4 (1): 1–38.
López Meirama, B. 1997. *La posición del sujeto en la cláusula monoactancial del español. Lalia*, Series Maior 7. Santiago de Compostela: Universidade de Santiago de Compostela.
Melis, Ch. 2011. “Los verbos de suficiencia.” *Lingüística mexicana* 6 (2): 29–59.
Pedersen, J. 2016. “Spanish Constructions of Directed Motion: A Quantitative Study.” En *Corpus-based Approaches to Construction Grammar*, eds. J. Yoon y S. Gries, 105–144. Amsterdam: John Benjamins.
Recalde, M. y V. Vázquez Rozas. 2009. “Problemas metodológicos en la formación de corpus orales.” En *A Survey of Corpus-Based Research*, eds. P. Cantos Gómez y A. Sánchez Pérez, 51–64. Murcia: Asociación Española de Lingüística de Corpus.
Rodríguez Espiñeira, M.J. 2000. “Percepción directa e indirecta en español: diferencias semánticas y formales,” *Verba* 27: 33–85.
Rodríguez Espiñeira, M.J. 2006. “Esquemas sintácticos con predicados cognitivos y predicativos obligatorios.” *Signo y Seña* 15: 113–138.
Rojo, G. 2001. “La explotación de la base de datos sintácticos del español actual.” En *Lingüística con corpus*, ed. J. de Kock, 255–286. Salamanca: Universidad de Salamanca.
Rojo, G. 2011. “Sobre la frecuencia de verbos y esquemas sintácticos.” En *Sintaxis y análisis del discurso hablado en español: homenaje a Antonio Narbona*, eds. J.J. de Bustos Tovar, R. Cano-Aguilar, E. Méndez García de Paredes y A. López Serena, 907–922. Sevilla: Universidad de Sevilla.
Rojo, G. 2014. “Hispanic corpus linguistics.” En *The Routledge Handbook of Hispanic Applied Linguistics*, ed. M. Lacorte, 371–387. Londres y Nueva York: Routledge.
Rojo, G. 2016. “Los corpus textuales en español.” En *Enciclopedia lingüística hispánica*, ed. J. Gutiérrez-Rexach, 285–296. Londres y Nueva York: Routledge.
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
Travis, C.E. y R. Torres Cacoullos. 2012. “Discourse Syntax.” En *The Handbook of Hispanic Linguistics*, eds. J.I. Hualde, A. Olarrea y E. O’Rourke, 653–672. Oxford: Wiley-Blackwell.
Vaamonde, G. 2011. *La alternancia posesiva con nombres de partes del cuerpo. Un estudio descriptivo del español a partir de datos de corpus*. Tesis doctoral. Universidade de Vigo.
Vaamonde, G., F. González Domínguez y J.M. García-Miguel. 2010. “ADESSE, a Database with Syntactic and Semantic Annotation of a Corpus of Spanish.” En *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC’10), 1903–1910. http://www.lrec-conf.org/proceedings/lrec2010/pdf/859_Paper.pdf
Vanderschueren, C. 2013. *Infinitivo y sujeto en portugués y español: un estudio empírico de los infinitivos adverbiales con sujeto explícito*. Berlín: Walter de Gruyter.
Vázquez Rozas, V. 2006. “*Gustar* Type Verbs.” En *Functional Approaches to Spanish Syntax: Lexical Semantics, Discourse and Transitivity*, eds. J.C. Clements y J. Yoon, 80–114. Basingstoke: Palgrave Macmillan.
Vázquez Rozas, V. y V. Miglio. 2016. “Constructions with Subject vs. Object Experiencers in Spanish and Italian: A Corpus-based Approach.” En *Corpus-based Approaches to Construction Grammar*, eds. J. Yoon y S. Gries, 65–102. Amsterdam: John Benjamins.

Yoon, J. y S. Wulff. 2016. “A corpus-based study of infinitival and sentential complement constructions in Spanish.” En *Corpus-based Approaches to Construction Grammar*, eds. J. Yoon y S. Gries, 145–164. Amsterdam: John Benjamins.