Lingüística de corpus: de los datos textuales a la teoría lingüística

José M. García-Miguel
Universidade de Vigo
gallego@uvigo.es

Resumen

Este artículo es una presentación general de la lingüística de corpus en el que se expone qué es un corpus lingüístico, qué relación tiene con otros tipos de datos, por qué es necesario anotarlo y cómo es el proceso de anotación. También se pasa revista a algunas de las tareas más comunes en la investigación lingüística basada en corpus, tales como la obtención de listados de frecuencias, la exploración de concordancias o la búsqueda de coapariciones (colocaciones) y otros tipos de información contextual. A lo largo del texto se intenta mostrar la relevancia de este tipo de datos para la teoría lingüística, en particular, para los modelos basados en el uso, como los cognitivos y funcionales.

Palabras clave: corpus, anotación, datos, frecuencia, coapariciones, teoría lingüística.

Abstract

In this paper a general presentation of Corpus Linguistics is provided by explaining what a linguistic corpus is, how it is related to other types of data, why it is necessary to annotate it, and what the annotation process is like. Some of the more common tasks in corpus-based linguistic research are also reviewed, such as obtaining frequency lists, exploring concordances, or finding co-occurrences (collocations) and other types of contextual information. Throughout the text, an attempt is made to show the relevance of this type of data for linguistic theory, in particular for use-based models, such as the cognitive and functional ones.

Keywords: Corpus, annotation, data, frequency, co-occurrences, linguistic theory.

1. Introducción: ¿qué es la lingüística de corpus?

La lingüística de corpus es un conjunto de metodologías relacionadas con la compilación y explotación de corpus en los estudios lingüísticos tanto teóricos como aplicados. La característica distintiva de esta rama de la lingüística es pues la utilización de corpus textuales como fuente primaria de datos. Un corpus puede definirse (por ejemplo, Sinclair 1991: 171; McEnery et al. 2006: 5; Gries 2009: 7; Rojo 2021: 1) como una colección de textos orales o escritos producidos en un contexto comunicativo natural, representativos de una lengua o variedad de lengua, almacenados en soporte informático y destinados al análisis lingüístico.

Los textos que componen un corpus (por ejemplo, una obra literaria, una noticia de periódico o una conversación entre amigos) fueron producidos con la intención de comunicar algo en cierto contexto y no generados más o menos artificialmente para investigar o ilustrar algún fenómeno lingüístico. Por tanto, los lingüistas de corpus basan sus descripciones y teorías en el uso real del lenguaje por parte de los hablantes, y en eso

ISSN: 1139-8736 https://infoling.org/elies/

se contrapone a otros métodos de obtención de datos tales como la introspección, las encuestas o la experimentación. En apartados posteriores volveremos sobre este punto y sus repercusiones teóricas.

Desde hace mucho tiempo se han hecho recopilaciones de textos con diferentes propósitos, incluido el análisis lingüístico; pero para llegar el estado actual de la disciplina fue crucial el desarrollo y difusión de la informática con su capacidad para almacenar y procesar datos lingüísticos. Por tanto, actualmente un corpus no es una simple colección de textos, sino un conjunto de textos digitalizados, que se pueden procesar mecánicamente. Toda la lingüística de corpus actual entra dentro de lo que Leech (1992: 106) prefiere etiquetar como "computer corpus linguistics".

Sin embargo, la lingüística de corpus, tal como la entendemos actualmente, ya tiene una tradición de más de medio siglo, que podemos remontar hasta la aparición en 1964 del primer corpus de la era moderna, el corpus Brown, que está conformado por 500 fragmentos de textos de inglés americano de un tamaño de aproximadamente 2000 palabras cada uno, hasta totalizar un millón de palabras. Tras estos primeros corpus, fueron apareciendo, no solo para el inglés sino también para otras lenguas europeas, corpus de referencia que pretendían ser muestras equilibradas y representativas de toda la lengua (tanto en sincronía como en diacronía). Un ejemplo prototípico de corpus es el *British National Corpus* (BNC), que contiene muestras de inglés británico hablado (10 %) y escrito (90 %) que suman un total de 100 millones de palabras.

En el caso del español, tenemos a nuestra disposición los corpus de la RAE (CREA, CORDE, CORPES) o la versión histórica del *Corpus del español* [CdE-hist] de M. Davies. De ellos, CORDE y CdE-hist son corpus diacrónicos que recogen textos desde los orígenes del español hasta el siglo XX, mientras que CREA y CORPES son corpus del español contemporáneo, que recogen en distintas proporciones textos tanto orales como escritos procedentes de España y América, de diferentes géneros y ámbitos temáticos. Cada uno de ellos contiene entre 100 y 400 millones de palabras.

Actualmente, la disponibilidad de textos que nos proporciona internet ha facilitado la creación de corpus de mucho mayor tamaño por el procedimiento de descargar textos de manera bastante indiscriminada a costa de sacrificar en mayor o menor medida el equilibrio, la representatividad y la diversidad de la muestra (para saber más sobre *big data*, véase Valenzuela 2022). La plataforma de *Corpus del español* dispone de un corpus (CdE-web) de 2 mil millones de palabras y otro de noticias (CdE-NOW) con más de 5 mil millones. En la plataforma *Sketch Engine* se pueden consultar corpus del español (esTenTen18) y de otras lenguas de un tamaño cercano a los 20 mil millones de palabras, llegando al doble en el caso del inglés (enTenTen20).

Existen muchos corpus de diferentes tipos, tamaños y composiciones (puede verse una tipología detallada en Torruella 2017 y un panorama general de corpus del español en Rojo 2016). Junto a los corpus de referencia representativos de una lengua, como el CREA o el CORPES, y los corpus diacrónicos como el CORDE, existen corpus que se centran en alguna variedad en particular. Por ejemplo, hay corpus de lenguaje especializado (textos jurídicos, informáticos, económicos, etc.), como el *Corpus Técnico do Galego*, o el *Corpus Tècnic del IULA* (multilingüe). Los corpus multilingües, como este último, pueden tener textos obtenidos por procedimientos similares que faciliten la comparación entre lenguas (*corpus comparables*, como la familia de corpus TenTen), o consistir en traducciones de una lengua a otra(s) convenientemente alineadas (*corpus*

ISSN: 1139-8736 https://infoling.org/elies/

paralelos, como CLUVI). También hay corpus de textos producidos por personas de determinado rango de edad, como los diferentes corpus de habla infantil contenidos en CHILDES o como COLA, que contiene lenguaje de adolescentes. Así mismo hay corpus de aprendices, como CAES o CORELE, que contienen textos producidos por quienes están aprendiendo una lengua y no por hablantes nativos. Y también existen otros muchos corpus compilados con propósitos y contenidos específicos. Mención especial merecen los corpus orales, como CORLEC, ESLORA, Val.Es.Co o los corpus del proyecto PRESEEA, dadas las dificultades de la recogida de datos orales tanto en la fase de grabación como en la de transcripción. Mayores aún son las dificultades para compilar corpus de lenguas de signos, aunque, en los últimos años, estamos viendo aparecer corpus de varias lenguas signadas como Auslan, BSL, NGT, DGS, o más recientemente CorLSE. Cada corpus nos proporcionará datos para estudiar en profundidad cierta(s) variedad(es) de lengua(s) y no otras; por lo que para quien utilice esta metodología siempre será importante seleccionar un corpus apropiado para los objetivos de su investigación.

A los corpus existentes y disponibles podemos añadir la posibilidad de que un lingüista individual pueda construir su propio corpus para atender a los objetivos de su investigación. En cualquier ordenador podemos disponer de herramientas, muchas de ellas gratuitas, que nos sirven de ayuda para recoger textos para compilar un corpus, añadirle automáticamente anotación lingüística y examinar los textos con herramientas típicas de la lingüística de corpus: concordancias, agrupaciones, listas de palabras, palabras clave, etc. Además, casi todo ordenador personal suele tener instalados una serie de programas de uso habitual que pueden ser muy útiles para trabajar con corpus, tales como editores de texto, hojas de cálculo, gestores de bases de datos, programas estadísticos, etc. Un plus bastante recomendable para quien se anime a trabajar con su propio corpus personal es el de tener conocimientos básicos de programación; pero, incluso con escasos o nulos conocimientos de programación, cualquier lingüista puede actualmente hacer lingüística de corpus con su ordenador personal, bien sea manejando sus propios datos o bien consultando corpus existentes elaborados por otras personas o por grandes equipos de trabajo.

En las siguientes secciones se ofrecerá un panorama general, inevitablemente algo superficial, de la metodología lingüística basada en corpus tomando como principal punto de vista el de un usuario lingüista. En la sección 2 se sitúan los datos tomados de corpus en relación con otras fuentes de datos lingüísticos. En la sección 3 se observa el proceso de compilación de corpus desde la selección de textos hasta su anotación y puesta a disposición de los usuarios. En la sección 4 se detallan algunas de las operaciones más básicas que un lingüista puede hacer con un corpus: buscar y obtener muestras de uso contextualizado de unidades lingüísticas, contar frecuencias de uso y documentar y computar coapariciones en el contexto inmediato. En la sección 5 se exponen brevemente las conexiones de la lingüística de corpus con algunas tendencias teóricas contemporáneas.

2. La lingüística de corpus y las fuentes de datos para hacer lingüística

Como en toda ciencia, las descripciones y teorías lingüísticas deben estar basadas en datos empíricos. Con carácter general, en lingüística podemos distinguir entre datos basados en la observación del uso del lenguaje por parte de los hablantes y datos basados en los

ISSN: 1139-8736 https://infoling.org/elies/

juicios y reacciones de los hablantes sobre la lengua, tales como los proporcionados por las encuestas o por los métodos experimentales. En este último grupo podría incluirse también la introspección, los juicios que el propio lingüista hace sobre su lengua. Fillmore (1992: 35) traza una conocida caricatura de dos especies de lingüistas según el tipo de datos que utilizan: por un lado, el "lingüista de sillón" que elabora teorías lingüísticas fiándolo todo a la introspección y sin apenas otros datos sobre los que sustentarlas y, por otro, el "lingüista de corpus" que acumula datos y más datos sin llegar a decir nada interesante sobre ellos. El primero ve cuestionada la validez empírica de sus propuestas, el segundo de los dos ve cuestionado el interés teórico de sus observaciones. Es posible que la polarización entre estos dos tipos de lingüistas se tome a veces demasiado en serio (Gries 2010b: 331), pero la descripción humorística de Fillmore nos recuerda la dialéctica entre reflexión teórica y recopilación de datos, que debe darse en toda ciencia y que apunta al mismo tiempo al hecho de que hay múltiples maneras de obtener datos para hacer lingüística. Más en serio, Fillmore observa que ningún corpus, por grande que sea, contiene todo lo que un lingüista puede llegar a necesitar; pero que, en todo corpus, aunque sea pequeño, podemos encontrar hechos que no se encuentran de otra manera. Su conclusión es que los dos tipos de lingüistas se necesitan mutuamente, o mejor aún, que los dos tipos de lingüistas deberían coexistir, siempre que sea posible, dentro del mismo cuerpo (Fillmore 1992: 35). La recomendación general es, pues, que la reflexión teórica vaya acompañada de datos (y viceversa), y utilizar diferentes metodologías de obtención de datos, muchas veces combinándolas o contrastándolas, y, en todo caso, siempre recurrir a la metodología que mejor se adapte a los objetivos de nuestra investigación.

La lingüística de corpus se caracteriza porque toma como fuente primaria de datos lo que los hablantes efectivamente dicen y escriben cuando se comunican por medio del lenguaje. Gilquin y Gries (2009: 5) ordenan los tipos de datos lingüísticos en una escala de más natural a menos natural, situando en un extremo los corpus de textos escritos y en el otro extremo los métodos experimentales en los que se pide a los participantes que hagan cosas con el lenguaje que normalmente no hacen (tales como reaccionar en condiciones de laboratorio ante ciertos estímulos lingüísticos o visuales). Frente a otros métodos de obtención de datos, la lingüística de corpus trata con datos naturales en cuya producción no ha influido el analista. Esto es, los textos que componen un corpus (sean obras literarias o científicas, artículos de prensa, páginas web, folletos, manuales de instrucciones, cartas…) se produjeron en su momento con un propósito que no era el de formar parte de un corpus. Con todo, debido a lo que se conoce desde Labov (1972: 171) como la *paradoja del observador*, algunos contextos comunicativos y algunas variedades lingüísticas son difíciles de observar sin influir en la producción. Por eso algunos corpus incluyen también transcripciones de conversaciones en las que participa el analista (observador participante), como ocurre en entrevistas sociolingüísticas, o en producciones discursivas que se obtienen como reacción a ciertos estímulos. En cualquier caso, para que podamos hablar de un corpus este debe contener textos (no necesariamente completos), puesto que un conjunto de palabras u oraciones aisladas, incluso aunque estas se hayan documentado en contextos naturales reales, no constituyen un corpus.

Por todo ello, la lingüística de corpus puede verse como una aproximación empírica cuyo punto de partida son los datos de uso del lenguaje tal como se realiza en los textos. Pero en tanto que aproximación que ha desarrollado sus propios métodos, la lingüística de corpus es muy diferente del análisis del discurso, pues su objetivo no es el análisis e interpretación de instancias individuales de uso, sino encontrar en la colección de textos

ISSN: 1139-8736 https://infoling.org/elies/

que constituye el corpus patrones regulares de uso que nos ayuden a entender la estructura y funcionamiento de un sistema lingüístico o a entender prácticas sociales, actitudes e ideologías que se reflejan en los usos lingüísticos. Tognini-Bonelli (2001) resume, como se muestra en la Tabla 1, las diferencias entre análisis textual y lingüística de corpus.

Texto	**Corpus**
Se lee entero	Se lee fragmentado
Se lee horizontalmente	Se lee verticalmente
Se lee el contenido	Se buscan patrones formales
Se lee como evento único	Se buscan eventos repetidos
Se lee como acto individual	Se estudia como muestra de una práctica social
Evento comunicativo coherente	No es un evento comunicativo coherente
Instancia de actuación individual (*parole*)	Ayuda a entender el sistema lingüístico (*langue)*

Tabla 1. Análisis de texto y análisis de corpus (adaptado de Tognini-Bonelli 2001: 3)

Los métodos de la lingüística de corpus también presentan diferencias con los utilizados para documentación de lenguas y variedades lingüísticas poco descritas. Para la mayoría de las lenguas es imposible obtener datos por el simple procedimiento de descargar textos de internet. En las etapas iniciales de descripción, necesitamos recoger léxico, patrones gramaticales y textos normalmente mediante procesos de elicitación con hablantes nativos. La recogida de textos puede llegar a constituir un corpus representativo útil para los propósitos de la descripción. Lo esperable y deseable es que cada vez tengamos más datos y más corpus sobre más lenguas. Sin embargo, la lingüística de corpus no se ha desarrollado históricamente como método para documentar lenguas de las que no teníamos datos, sino como conjunto de herramientas para el análisis de lenguas, como el inglés o el español, para las que existe una larga tradición descriptiva apoyada en una larga tradición escrita y, por tanto, en la disponibilidad de textos.

Lo que ha cambiado frente a la lingüística de hace un siglo es que la tecnología actual nos permite manejar en poco tiempo una cantidad de datos antes impensable. Y eso es lo que nos ha traído la informática. El desarrollo de los corpus como herramienta para los lingüistas se ha comparado con la invención del telescopio, que permitió observar cosas que no se habían visto nunca antes (Stubbs 1996: 231 y, tras él, muchos otros). El telescopio aproxima el objeto de estudio, permite una observación más precisa y, con ello, la formulación de teorías más coherentes y comprehensivas sobre el universo. La revolución informática moderna permite el rápido procesamiento de ingentes cantidades de datos, lo que ha revolucionado todas las ciencias, incluida la lingüística. El físico Freeman Dyson (1997: 50-51) explica que en la historia de la ciencia pueden producirse revoluciones conceptuales [‘*concept-driven revolutions’*], las cuales explican cosas viejas de maneras nuevas, y revoluciones instrumentales [‘*tool-driven revolutions’*], las cuales descubren cosas nuevas que hay que explicar, y añade que las revoluciones instrumentales son más frecuentes que las conceptuales. Pone como ejemplos el impacto que tuvieron en su tiempo el telescopio, luego el microscopio, y más recientemente los computadores. La lingüística de corpus forma parte de una revolución instrumental en la que la informática pone a nuestra disposición muchos datos nuevos que debemos explicar.

Una propiedad destacable de la lingüística de corpus es la exhaustividad, entendida como la obligación de intentar un examen de todos los casos contenidos en el corpus. El análisis y las interpretaciones teóricas deben dar cuenta de todos los datos, sin hacer una selección previa que descarte aquellos casos que decidimos ignorar por irrelevantes. Este principio

ISSN: 1139-8736 https://infoling.org/elies/

de exhaustividad en relación con los datos disponibles es un destacable punto fuerte de la lingüística de corpus (Leech 1992: 112). Cualquier corpus nos va a presentar ejemplos abundantes de fenómenos muy comunes, evitando que nuestras descripciones se basen en fenómenos llamativos que atraen nuestra atención y minusvaloren lo más común. Pero el principio de exhaustividad lleva a que cualquier corpus nos pueda presentar muchos casos de fenómenos 'raros', poco frecuentes, y nos obligue a preguntarnos si nuestras categorías analíticas valen para todos los casos posibles.

3. De los textos a los corpus anotados

El lingüista que quiere utilizar en sus investigaciones datos de corpus debe tener en cuenta que tanto la compilación como la utilización o explotación de corpus conllevan en todas sus fases una serie de problemas prácticos que pueden o suelen tener carga teórica: ¿cómo seleccionamos una muestra de textos que sea representativa y equilibrada?, ¿cómo digitalizamos los textos escritos?, ¿cómo transcribimos los textos orales y signados?, ¿cómo reconstruimos la relación entre el texto-producido y el contexto dinámico en que se produjo?, ¿cómo identificamos y clasificamos las unidades básicas (palabras, oraciones…)? Analizaremos brevemente en esta sección algunos de estos problemas desde el punto de vista del lingüista usuario de esta metodología, teniendo en cuenta también el punto de vista de los creadores de corpus.

El primer problema al que nos enfrentamos es la selección de los textos que constituyen el corpus o, para un usuario final, la selección de un corpus u otro. Por grande que sea, ningún corpus contiene todos los discursos producidos en una lengua y solo en casos muy específicos podría llegar a contener todos los textos existentes de cierta variedad lingüística. Los corpus consisten en muestras que pretenden ser representativas del conjunto de la lengua y, por eso, siempre que se diseña un corpus deben considerarse con atención los criterios de selección (véase Biber 1993; Toruella y Llisterri 1999). Sin embargo, como muy raramente conocemos el tamaño y composición exacta de la población total de la que se obtiene la muestra (el conjunto total de textos orales y escritos producidos en una variedad lingüística) ni son igualmente accesibles todos los tipos de discurso, la representatividad y el equilibrio de un corpus es solo un ideal teórico prácticamente imposible de alcanzar (véase Stefanowitsch 2020: 29; Rojo 2021: 291-294). En cambio, una vez decidida la lengua o variedad de lengua, se intenta que la muestra sea diversificada, que contenga textos de diferentes géneros, temáticas, procedencias geográficas, etc. de manera que ningún tipo particular tenga un peso excesivo que pueda sesgar los resultados que obtengamos. Esa es también una de las razones por las que, en los primeros corpus, como el de Brown, se incluían fragmentos textuales no demasiado grandes y no obras completas. En los corpus modernos de gran tamaño y diversificación es más difícil que una obra o género particular tenga un peso excesivo. En cualquier caso, es muy importante que al extraer conclusiones de un corpus tengamos en cuenta de qué es representativo ese corpus y qué carencias o sesgos puede tener, antes de pretender generalizar a toda una lengua lo que quizá sea específico de esa muestra.

Por otro lado, aunque las diferentes secciones de un corpus no tengan tamaños proporcionales a los de la población total, lo que nos interesa habitualmente no son las frecuencias absolutas de un fenómeno, sino frecuencias relativas o normalizadas que nos

ISSN: 1139-8736 https://infoling.org/elies/

permitan comparar una sección con otra independientemente del tamaño de cada muestra. Si el corpus es grande podríamos obtener datos suficientes sobre el fenómeno de interés y observar a posteriori cómo se distribuyen en las diferentes secciones del corpus para determinar en qué variedades lingüísticas ocurren esos datos. Para el lingüista que hace investigación basada en corpus es crucial conocer cómo está hecho el corpus que utiliza y saber cuáles fueron los criterios de selección de textos y qué sesgos puede tener esa selección. De eso depende en buena medida la validez de las conclusiones que obtenga. Casi más importante aún que el equilibrio de la muestra es que los textos que componen el corpus estén catalogados con información sobre las circunstancias de su producción (género, medio, ámbito temático, fecha, lugar de origen…) o sobre las características de quienes los produjeron (sexo, edad, nivel sociocultural…). Estos metadatos son los que nos permitirán seleccionar subcorpus o realizar búsquedas con diferentes criterios y comparar unas variedades con otras.

Los textos que componen un corpus deben compilarse necesariamente en soporte informático para poder ser analizados con los métodos de la lingüística de corpus. La informatización es un problema relativamente complejo si queremos contar con muestras de discurso oral (que primero deben transcribirse) o con muestras de discurso escrito que hasta hace poco estaban mayoritariamente en soporte papel. Por el contrario, parece un problema trivial (aunque no tanto como parece a primera vista) si construimos un corpus recurriendo a las ingentes cantidades de textos que se producen actualmente en internet. Pero, sea cual sea el soporte original, debemos guardarlos en un formato que permita al usuario la consulta desde cualquier sistema informático utilizando las herramientas que proporciona la lingüística de corpus. En general, esto requiere registrarlos como texto plano sin ningún tipo de formato ni de elemento adicional. Un archivo de texto simple digitalizado no es ni más ni menos que una cadena de códigos que simbolizan caracteres (alfabéticos, numéricos, signos de puntuación y otros). Los textos escritos pueden codificarse directamente, pero hay que prestar atención al sistema de codificación, para evitar problemas con diacríticos y otros símbolos que van más allá del alfabeto latino básico. Lo más recomendable actualmente es el sistema Unicode, válido para cualquier sistema de escritura.

En este sentido, los textos orales, multimedia y signados pueden grabarse; pero, para poder ser utilizados en lingüística de corpus, deben transcribirse a texto escrito utilizando un sistema de transcripción apropiado. Dependiendo de los objetivos con los que se compila el corpus, podría ser suficiente con una transcripción ortográfica, que luego codificamos en un archivo de texto plano igual que cualquier otro texto escrito; pero muy probablemente también nos interese poder recuperar aspectos de la pronunciación, la prosodia y otros aspectos del discurso oral, tales como pausas, vacilaciones y solapamientos, que no se recogen en la escritura ortográfica convencional. Por su parte, en el caso de las lenguas de signos se añade el problema de que no existe un sistema estándar de signoescritura; por lo que, para poder explorar un corpus de una lengua de signos, lo habitual es acompañar las grabaciones en vídeo de transcripciones basadas en un sistema de glosas identificativas (‘id-glosas’) de cada seña (Johnston 2010; Pérez et al. 2019). Más allá de la simple transcripción o glosado de lenguaje oral o signado, uno de los grandes retos para la lingüística de corpus actual es la recopilación, codificación y explotación de corpus multimodales, consistentes en colecciones de grabaciones audiovisuales que sirven para estudiar la interacción entre las diferentes modalidades (voz, gestos, mirada, orientación corporal, proxémica, etc.) utilizadas simultáneamente

ISSN: 1139-8736 https://infoling.org/elies/

en la comunicación humana y la interacción entre ellas (Kipp et al. 2009; Knight y Adolphs 2020). Para su análisis necesitaremos varios niveles de transcripción y anotación alineados temporalmente con las grabaciones de audio y video. Herramientas como ELAN pueden ayudar en esa tarea.

Los textos escritos también suelen contener en su versión original mucha información que se pierde al digitalizarlos como texto plano: disposición espacial, tipografía, abreviaturas, gráficos, tablas, llamadas a notas, etc. Esa información puede ser relevante para la interpretación del texto y en el diseño de un corpus se debe determinar cuánta de esa información debe codificarse y con qué procedimiento. Lo mismo puede decirse de la estructura del texto, por ejemplo, la de un libro en capítulos y secciones, o la de una noticia de periódico en titular, entradilla y cuerpo del texto. Por tanto, la compilación de un corpus requiere, además de la codificación del texto como cadena de caracteres, la codificación de información externa sobre cada texto (los metadatos) o información interna sobre propiedades que acompañan al texto (paralenguaje, estructura interna, etc.). La mayoría de los corpus existentes actualmente contienen anotaciones de alguno de los siguientes tipos:

- Información que cataloga los textos según las circunstancias de su producción (género, medio, ámbito temático, fecha, lugar de origen…) o según las características de quienes los produjeron (sexo, edad, nivel sociocultural…)
- Información sobre rasgos paralingüísticos que acompañan al texto, tales como el tipo y tamaño de fuente en los textos escritos o la prosodia en los textos orales transcritos. También puede incluirse información o referencias sobre elementos simbólicos multimodales, tales como gráficos que acompañan a textos escritos o gestos con textos orales.
- Información sobre la estructura del texto: capítulos, secciones, secuencias temáticas, movimientos conversacionales, …

Esta codificación recibe a veces el nombre de *anotación*, pero normalmente se reserva este término o el de *etiquetación* para la codificación de propiedades lingüísticas relativamente abstractas que no pueden observarse directamente en el texto. Si un usuario quiere recuperar datos de un corpus consistente en archivos de texto podrá buscar con el software apropiado cadenas de caracteres que tal vez correspondan a (partes de) palabras y frases. Podemos obtener así rápidamente los datos buscados, pero difícilmente podemos encontrar de esta manera propiedades abstractas como, por ejemplo, una estructura sintáctica particular. Para ello sería necesario anotar el texto, esto es, añadir indicaciones explícitas de propiedades lingüísticas de los elementos del texto, tales como clase de palabra, lema, acepción de una palabra polisémica, o la estructura sintáctica o semántica de una oración. Cualquier nivel de análisis lingüístico es susceptible de anotación. Es posible, por tanto, que un corpus contenga anotación de diversos tipos:

- Fonológica: transcripción fonética, límites silábicos, rasgos prosódicos (acento, tono, entonación)
- Ortográfica: asociación de variantes ortográficas con formas normalizadas
- Morfológica: raíces, prefijos, sufijos…
- Léxico-gramatical: lematización, clase de palabra (N, V), rasgos morfosintácticos (singular, plural, tiempo verbal, …)

ISSN: 1139-8736 https://infoling.org/elies/

- Sintáctica: constituyentes, relaciones de dependencia, funciones sintácticas (Suj, Obj), categorías sintácticas...
- Semántica: desambiguación léxica, campos semánticos/ontologías, papeles sintáctico-semánticos, clasificación de nombres propios, metáforas, …
- Textual-discursiva: relaciones anafóricas, tema/rema, información dada/nueva, estructura del discurso, …
- Pragmática: actos de habla, análisis de sentimientos y opiniones, roles discursivos, conocimiento contextual compartido, …

De estos niveles de anotación, el más fácil de implementar y por ello el más extendido con mucha diferencia en los corpus disponibles para uso general es la anotación léxico-gramatical consistente en la lematización y etiquetado morfosintáctico de las palabras del corpus. Con la lematización adscribimos cada palabra ortográfica a una forma de diccionario o lema. En el etiquetado morfosintáctico se incluye, al menos, la clase de palabra y usualmente también las categorías flexivas. Por ejemplo, la palabra ortográfica *cuento* puede corresponder al singular del sustantivo *cuento* o a la primera persona singular del presente de indicativo del verbo *contar*.

Toda anotación implica siempre una interpretación del texto, por lo que requiere siempre un proceso de toma de decisiones informado teóricamente, aunque condicionado por limitaciones prácticas. Y esto vale para cualquier nivel de análisis, incluso los aparentemente más triviales. Por ejemplo, la tokenización o división del texto en unidades elementales susceptibles de recibir anotaciones ulteriores suele consistir en la división del texto en palabras ortográficas delimitadas por espacios y signos de puntuación. Aunque parece un proceso sencillo, inmediatamente nos encontramos con problemas prácticos, pues algunos signos de puntuación no separan palabras ortográficas: los guiones unen a veces palabras compuestas y los puntos podemos encontrarlos en el interior de siglas y abreviaturas. Más importante aún es el hecho de que el estatus teórico de la palabra como unidad básica del análisis lingüístico es discutible y que los límites entre palabras son indeterminados (Haspelmath 2011). En español, por ejemplo, podemos discutir el estatus como palabra de clíticos pronominales (que a veces se unen ortográficamente al verbo y otras no), pero también el estatus de artículos, preposiciones, conjunciones y otras formas dependientes.

El proceso de anotación en lingüística de corpus acarrea diversos problemas prácticos relacionados con cómo se realiza y con cómo se codifica. Otras facetas del proceso de anotación tienen que ver con el diseño de esquemas de anotación, desarrollo de estándares, evaluación y usos de las anotaciones, etc. (Ide y Pustejovky 2017).

En cuanto al procedimiento, la anotación puede ser manual o automática. La anotación manual solo es factible con corpus pequeños y, aun así, es muy trabajosa y está sujeta a imprecisiones e inconsistencias. La anotación manual es compatible con cualquier punto de vista teórico y permite a uno decidir qué es lo que quiere anotar, según las necesidades de su investigación. Ahora bien, se puede facilitar la comparación entre corpus anotados si se utilizan los mismos esquemas de anotación, de ahí la conveniencia de desarrollar estándares de anotación. En cualquier caso, se requiere una definición previa de las categorías que se van a utilizar en la anotación y de los criterios de delimitación. Si los criterios están bien definidos deberían tener validez intersubjetiva, esto es, personas diferentes llegarían a clasificaciones similares de los datos. Conviene, además, utilizar algún índice estadístico para cuantificar el grado de acuerdo entre anotadores diferentes.

ISSN: 1139-8736 https://infoling.org/elies/

Con corpus grandes solo es factible la anotación automática. Existen programas suficientemente eficaces, muchos de ellos gratuitos, para ciertos tipos de anotación. En la lematización y el etiquetado morfosintáctico de las palabras del corpus se alcanzan porcentajes de acierto superiores al 95 %. Sin embargo, un error inferior al 5 % hace que en un corpus de 100 millones de palabras tengamos casi 5 millones de palabras anotadas erróneamente y que en una oración elegida al azar una palabra de cada 20 esté mal etiquetada. Para otros niveles de análisis, como la anotación sintáctica o la desambiguación de significados léxicos, el porcentaje de acierto de los programas existentes es menor y por eso son mucho más raros los corpus disponibles con ese tipo de anotación. Recursos que contienen análisis sintáctico y semántico de corpus, tales como ADESSE y BDS, AnCorA, SenSem o el IULA Spanish Treebank, han sido elaborados a mano en su totalidad o en un porcentaje significativo, aunque lógicamente han contado con asistencia informática.

El proceso de anotación automática está condicionado, en primer lugar, como en la anotación manual, por la elección del propio esquema de anotación (por ejemplo, qué sistema de clasificación de palabras utilizar). Luego el programa categoriza los elementos mediante cálculo de probabilidades. Por ejemplo, en la frase *El cuento de los tres cerditos*, se trataría de calcular la probabilidad de que *cuento* sea nombre o verbo, junto con la probabilidad de que un nombre o un verbo aparezca tras el artículo *el* o ante la preposición *de*. No obstante, estas probabilidades, a su vez, suelen estar basadas en "corpus de entrenamiento" que han sido anotados o revisados manualmente. La eficacia de un proceso automático de anotación dependerá pues del esquema de anotación y de la cantidad y calidad del corpus de entrenamiento; y de si ha sido posible una ulterior revisión manual de la anotación automática.

Para el usuario de un corpus ya anotado, sea manual o automáticamente, es importante tener en cuenta que toda anotación es el resultado de un proceso de toma de decisiones llevado a cabo por otros investigadores (Stefanowitsch 2020: 83). Cualquier usuario de un corpus debe tomar la precaución de no asumir acríticamente las anotaciones que encuentra en el corpus y debe conocer los criterios que se han utilizado en la anotación. Y viceversa, quienes anoten un corpus deben explicitar los criterios de anotación. Siendo así, la anotación siempre añade valor a un corpus, incluso si contiene categorizaciones discutibles de los datos.

Existen diferentes maneras de codificar la anotación de un corpus. Un procedimiento común hace un tiempo es utilizar etiquetas unidas a la unidad etiquetada mediante un separador predefinido, como un guion bajo, como, por ejemplo, en *El*_ART *cuento*_NC *de*_PREP *los*_ART *tres*_NUM *cerditos*_NC. Con cualquier software de exploración de corpus (incluso con cualquier editor de texto) utilizando comodines o patrones de búsqueda con expresiones regulares podríamos buscar fácilmente elementos del texto, categorías anotadas o una combinación de ambas.

Mucho más potente y más extendido es el lenguaje de marcado XML que nos permite codificar virtualmente cualquier cosa utilizando etiquetas entre ángulos (por ejemplo, encerrando una oración entre <oración></oración>) y pares atributo = valor acompañando a las etiquetas (por ejemplo, gen="masc"). Este lenguaje puede servir para hacer anotaciones lingüísticas, como en el ejemplo (1), y también es el más utilizado para codificar los metadatos y la estructura de los documentos de un corpus (para lo cual, es recomendable seguir las recomendaciones de la "Text Enconding Initiative" [TEI]).

ISSN: 1139-8736 https://infoling.org/elies/

(1)

```
<oración id="1">
 <palabra lema="el" cat="artículo" gen="masc" num="singular" > El </palabra>
 <palabra lema="cuento" cat="nombre" gen ="masc" num="singular" > cuento
</palabra>
 <palabra lema="de" cat=" preposición"> de </palabra>
 <palabra lema="el" cat="artículo" gen="masc" num="plural" > los </palabra>
 <palabra lema="3" cat="numeral" > tres </palabra>
 <palabra lema="cerdo" cat="nombre " gen="masc" num="plural"> cerditos </palabra>
</oración>
```

Otra posibilidad de codificación, que puede ser bastante práctica desde diferentes puntos de vista, es utilizar algún tipo de formato tabular donde cada fila corresponde a un dato (puede ser una unidad lingüística particular, como una palabra o una oración) y cada columna a una categoría de clasificación (a una variable). Es el formato de salida estándar de analizadores automáticos como TreeTagger o FreeLing. En el ejemplo (2) se ha situado cada palabra en una fila y en las columnas se han anotado lema, clase palabra y relaciones sintácticas de dependencia. Un formato llamado CONLL-U, similar al de (2) pero con diez columnas, es el utilizado por universaldependencies.org, que contiene corpus de más de un centenar de lenguas analizadas sintácticamente con el mismo esquema de anotación.

(2)

ID	TOKEN	LEMMA	POS	HEAD	DEPREL
1	El	el	DET	2	det
2	cuento	cuento	NOUN	0	root
3	de	de	ADP	6	case
4	los	el	DET	6	det
5	tres	3	NUM	6	nummod
6	cerditos	cerdo	NOUN	2	nmod

El formato de filas y columnas también es adecuado para anotar a posteriori datos extraídos de corpus (por ejemplo, concordancias) añadiendo en columnas adicionales los rasgos que necesitemos para una investigación particular.

Combinando tablas de anotación, podemos construir una base de datos relacional, un formato que, según Davies (2005), presenta numerosas ventajas para la anotación de corpus, y que es lo que lo que utiliza el propio Davies en el trasfondo de las distintas versiones de *Corpus del Español* y otros corpus gestionados por él. Y es también, por ejemplo, el formato que se utiliza en BDS y ADESSE (García-Miguel et al. 2010).

Finalmente, el proceso de compilación, codificación y anotación de corpus debería culminar en un sistema de consulta y explotación. En los corpus puestos a disposición pública, lo importante no son tanto los formatos particulares utilizados para codificación y anotación del corpus como que los criterios de construcción del corpus estén claros y que el sistema de consulta nos permita explorarlo utilizando diferentes criterios de búsqueda. Un buen sistema de explotación de corpus nos permite refinar las búsquedas combinando criterios y, además, nos proporciona una clasificación inicial de los datos, que luego podrá refinar el usuario. Puede verse en De Benito Moreno (2019) un buen

ISSN: 1139-8736 https://infoling.org/elies/

análisis comparativo de las posibilidades que ofrece la interfaz web de trece corpus del español desde la perspectiva del usuario lingüista.

4. ¿Qué hacer con un corpus? Lo que (mejor) nos pueden ofrecer los corpus a los lingüistas

Con un corpus se pueden hacer múltiples cosas y se ha demostrado su utilidad casi en cualquier campo de los estudios del lenguaje. El desarrollo de la lingüística de corpus a partir de los años 70 fue impulsado por su utilidad para la lexicografía, la enseñanza de lenguas, la traducción y otros campos de la lingüística aplicada. Sin embargo, en este artículo y en este apartado nos centraremos en la relevancia de los datos de corpus para la lingüística teórica (en contraposición a la aplicada), esto es para la descripción y explicación de la estructura interna de las lenguas, en particular del léxico y la gramática. La lingüística de corpus, como estamos viendo, nos permite contar con datos naturales y nos proporciona herramientas para manejar una cantidad ingente de datos. Ahora bien, como toda herramienta y como toda fuente de datos, la lingüística de corpus nos inclina a fijarnos más en unas cosas que otras y a concebir el estudio del lenguaje de unas maneras mejor que de otras. Con carácter general, podemos decir que lo que mejor nos permite la lingüística de corpus es documentar si cierto fenómeno ocurre en el corpus, con qué frecuencia ocurre y cómo se distribuye en las diferentes secciones del corpus o en relación con otros elementos presentes en el corpus. Tomando la palabra (ortográfica) como unidad básica de referencia, cualquier sistema de exploración de corpus (sea un sistema de consulta online sea un sencillo programa de análisis textual como *AntConc*) nos permite obtener como mínimo alguno de los siguientes resultados:

- Listados de usos contextualizados de la palabra buscada (concordancias o ‘palabra clave en contexto’)
- Listados de las palabras del corpus (con su frecuencia)
- Listados de coapariciones: expresiones multipalabra, colocaciones, ...

Otros resultados que nos ofrecen algunos corpus, tales como palabras clave de un texto o conjunto de textos, tesauros, o resúmenes de la distribución de una palabra en el corpus, no son sino resultados derivados de los anteriores. En corpus anotados también podemos hacer lo mismo con elementos gramaticales, esto es buscar y obtener muestras de uso contextualizado y frecuencias de aparición y coaparición de cualquier categoría gramatical anotada y de patrones sintagmáticos, no solo de palabras individuales. Veremos a continuación con algo más de detalle los tipos básicos de explotación.

4.1. Documentar ocurrencias de uso

Para la persona que consulta un corpus, el procedimiento seguido en el análisis de corpus empieza normalmente por extraer de un corpus suficientemente grande fragmentos textuales que muestren fenómenos lingüísticos particulares y el objetivo deseable es que el sistema de consulta nos devuelva todos los casos y solo los casos del fenómeno investigado. En este sentido, uno de los usos más comunes de un corpus es la obtención de concordancias, que consisten en un listado de ocurrencias, presentado típicamente con una línea para cada ocurrencia, con la unidad buscada en el centro precedida y seguida

ISSN: 1139-8736 https://infoling.org/elies/

del contexto inmediato hasta cierto número de palabras. Además, encontraremos en las concordancias alguna referencia que identifica el ejemplo y, posiblemente, algún enlace que permita acceder a un contexto más amplio (a veces también a la propia grabación en el caso de corpus orales o multimedia). Por ejemplo, en (3) tenemos las primeras líneas resultantes de la búsqueda de ocurrencias de la palabra *ahorita* en Ameresco:

(3) *Ahorita* en corpus Ameresco

conversación	hablante	
HAV_046_02_17	B	pero es que dijiste **ahorita** que no ibas a ir
HAV_046_02_17	B	[pero] mira en esta misma premiación **ahorita** vamos a ponerlo
HAV_048_02_17	A	¿no fueron **ahorita**? a las ocho fueron ocho ocho y pico

En los casos más simples se buscarán cadenas de caracteres que se corresponden con (partes de) palabras o frases, con lo que podemos encontrar usos contextualizados de, por ejemplo, *ahorita*, *cuento*, *hermoso*, *democracia*, *alberca*, *delante mío*, etc. En consultas un poco más complejas, se pueden buscar patrones de cadenas de caracteres con comodines o lo que se conoce como *expresiones regulares* (es decir, una especie de fórmulas en las que ciertos símbolos pueden representar clases de caracteres). Por ejemplo, en muchos corpus buscando "*hermos*"* encontraremos las formas *hermoso*, *hermosa*, *hermosos*, *hermosas* y también *hermosura*, etc. Si el corpus está lematizado y etiquetado podremos buscar de la misma manera en las anotaciones correspondientes, por ejemplo, el sustantivo *cuento* (sin incluir formas del verbo *contar*), el adjetivo *hermoso* o la construcción formada por un adverbio seguido de un posesivo (tipo *delante mío/mía*). Y las búsquedas podremos hacerlas sobre todo el corpus o sobre una sección/subcorpus particular: por ejemplo, podemos buscar y comparar el uso de un taco como *joder*, en textos orales o escritos, en entrevistas frente a conversaciones, en producciones de hombres o de mujeres, en diferentes rangos de edad, etc.

Dependiendo de los niveles de anotación del corpus y del diseño del sistema de consulta, podremos refinar en mayor o menor grado las búsquedas y combinar diferentes criterios. Por ejemplo, la aplicación de consulta del CORPES, gracias al sistema de catalogación de documentos y a la lematización y etiquetación morfosintácticas, nos permite buscar ocurrencias del sustantivo *cuento* [lema] seguido de un adjetivo cualquiera [categoría del contexto] en textos orales procedentes de España [subcorpus]. Como resultado, obtenemos 41 casos en 33 documentos en la versión 0.94 de CORPES (Figura 1).

ISSN: 1139-8736 https://infoling.org/elies/

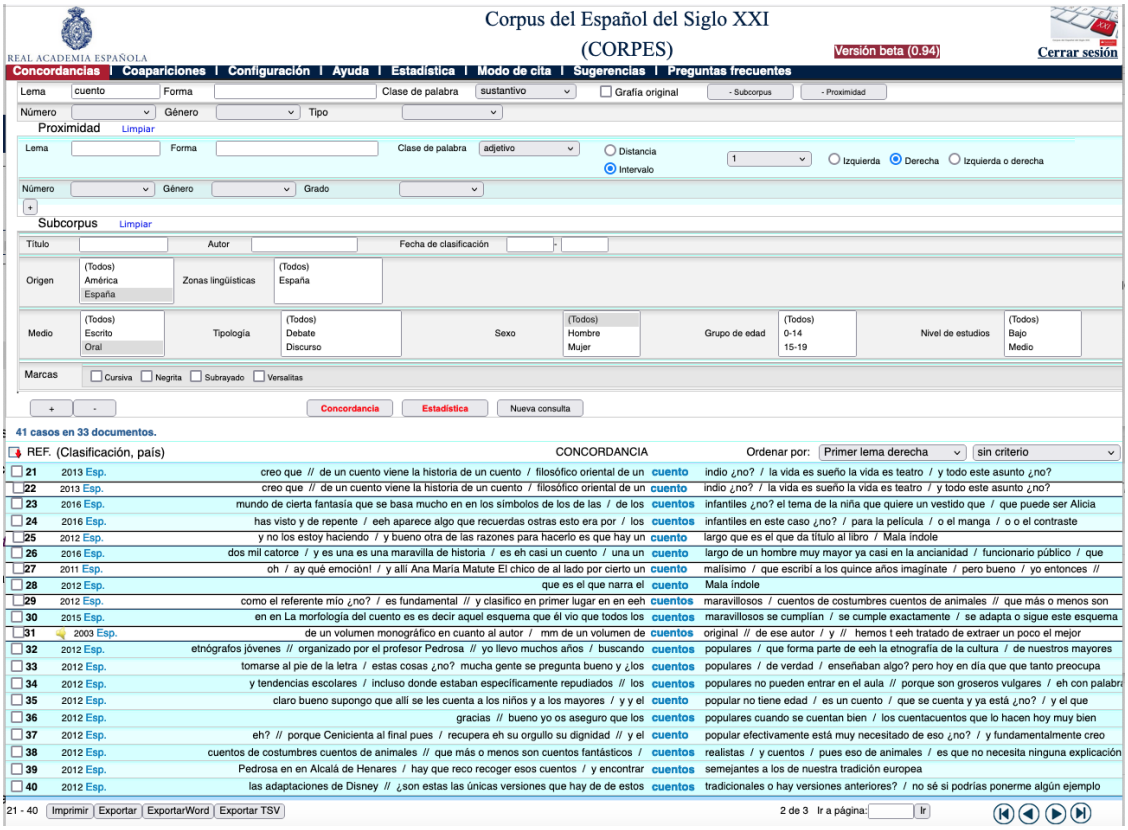

Figura 1. Concordancias del sustantivo *cuento* seguido de adjetivo en CORPES (subcorpus: medio oral y origen España)

El siguiente paso para el investigador consiste en observar los datos recogidos, ordenarlos y clasificarlos con diferentes criterios y elaborar hipótesis sobre las relaciones entre los datos observados. Es decir, se estudian las concordancias en busca de patrones regulares que se repiten en los ejemplos. Para ello suele ser muy útil la posibilidad de ordenar el listado por diferentes criterios, entre ellos el de ordenar por el pivote (esto es, las palabras ortográficas resultantes de la búsqueda) o por la primera palabra a la derecha o a la izquierda de la palabra buscada, de manera que aparezcan juntas algunas combinaciones que se repiten. En el ejemplo de la Figura 1 quizá fuera interesante ordenar por la primera palabra a la derecha para ver por orden alfabético qué adjetivos se registran en combinación con el sustantivo *cuento*. También podemos ordenar según alguno de los criterios de clasificación de los textos (fecha, autor, medio, etc.).

En el proceso de análisis lingüístico basado en corpus suele ser imprescindible un procesado manual de los ejemplos, es decir, una anotación de cada una de las ocurrencias utilizando criterios específicos de la investigación particular que se esté llevando a cabo. Por eso, nos interesa que el sistema de consulta permita la exportación de los resultados en un formato (.csv o .tsv) que sea fácilmente importable a una hoja de cálculo o a algún otro sistema de gestión de datos donde anotar en columnas adicionales las propiedades variables que atribuimos a cada instancia de uso.

Más arriba hemos mencionado que una propiedad destacable de la lingüística de corpus es la exhaustividad, lo que nos obliga a tener en cuenta todos los casos recogidos en el corpus según los criterios previamente establecidos y no hacer una selección de conveniencia. Puede ocurrir que estemos usando un corpus grande para investigar un

ISSN: 1139-8736 https://infoling.org/elies/

fenómeno muy frecuente, como suele ser el caso de la mayoría de los fenómenos gramaticales, y que sea imposible observar todos los ejemplos uno a uno. Pero, en general, lo que nos interesa en lingüística de corpus no suele ser el describir uno a uno todos los ejemplos encontrados sino encontrar tendencias cuantitativas generales. Para eso puede ser suficiente con la anotación que tenga el corpus. Si, por el contrario, debemos observar en los ejemplos del fenómeno analizado propiedades que no están previamente codificadas y la cantidad de ocurrencias es excesiva, lo que podemos hacer es tomar una muestra manejable que sea representativa del total. Algunos sistemas de consulta de corpus pueden proporcionar una selección aleatoria que permite sacar conclusiones sobre tendencias generales significativas.

4.2. Frecuencias

En la exploración de corpus, junto con las muestras de uso, que podemos analizar cualitativamente, obtendremos siempre datos de frecuencia. Se ha propuesto definir la lingüística de corpus como un conjunto de investigaciones que se pueden formular en términos de la distribución condicional de los fenómenos lingüísticos en un corpus (Stefanowitsch 2020: 56). De este modo, la lingüística de corpus se vincula necesariamente con los métodos cuantitativos, y con la estadística como disciplina que proporciona herramientas para el análisis cuantitativo. Como destaca Gries (2009: 11), en un corpus no hay ni significado ni funciones, solo cadenas de caracteres codificados informáticamente, por lo que lo único que realmente nos proporcionan los corpus son frecuencias de ocurrencia (cuánto aparecen ciertas palabras, construcciones gramaticales, etc. en un corpus particular) y frecuencias de coocurrencia (cuánto aparecen ciertas palabras en una construcción gramatical particular, etc.). Y le queda al investigador la tarea de interpretar esas frecuencias en términos significativos o funcionales.

Si lo único que podemos observar en el análisis de corpus son frecuencias de ocurrencias y coocurrencias en textos, lo relevante para nuestras descripciones lingüísticas es que al utilizar estos datos no nos estamos preguntando si algo es o no es posible en una lengua (menos aún, si es correcto o incorrecto), si acaso nos preguntaríamos si puede o no documentarse en el uso registrado en ese conjunto de textos. Por consiguiente, tampoco nos ponemos como objetivo la formulación de leyes que se cumplen necesariamente, en las que un fenómeno implica necesariamente otro. Lo que constatamos en el trabajo con corpus es que los elementos del lenguaje o sus combinaciones tienen una frecuencia variable y que la frecuencia relativa de cualquier unidad puede variar en relación con múltiples factores contextuales (registros, estilos, grupos sociales, dialectos geográficos, canal, diacronía, etc.) y también en relación con la coaparición de otras unidades textuales. En consecuencia, al tomar necesariamente en cuenta esa variación, la lingüística de corpus no opera con oposiciones discretas (sí o no), sino con gradaciones entre lo más o menos frecuente y, por tanto, necesita trabajar con métodos estadísticos.

Tomemos un ejemplo simple: lo que en España llamamos *ordenador* recibe en otros países hispanohablantes el nombre de *computadora*. Pero no se trata simplemente de que en la variedad *x* se utilice una forma y en la variedad *y* otra. En la Tabla 2. vemos los resultados globales que nos da el CORPES para seis grandes zonas hispanohablantes.

ISSN: 1139-8736 https://infoling.org/elies/

Zona	**ORDENADOR**		**COMPUTADORA**	
	Frec	Fnorm.	Frec	Fnorm.
España	10.636	87,86	549	4,53
México y Centroamérica	535	8,40	4.250	66,75
Río de la Plata	291	6,30	3.698	80,18
Andina	236	8,76	1.249	46,40
Antillas	236	10,48	942	41,85
Caribe continental	278	6,77	842	20,52

Tabla 2. Frecuencias absolutas y relativas (normalizadas por millón) de los lemas "ordenador" y "computadora" en el *Corpus del español del siglo XXI* (CORPES) v 0.94

Para analizar este ejemplo, debemos fijarnos en las frecuencias relativas o normalizadas (columna *Fnorm*), en este caso ocurrencias por cada millón de palabras, porque las frecuencias absolutas (columna *Frec*) lógicamente pueden depender del tamaño de cada sección del corpus. Con ello constatamos que, efectivamente, el uso de *ordenador* es mucho mayor en España que en otras zonas, mientras que el uso de *computadora* es mucho menor; pero que, en cualquier zona, podemos registrar usos de cualquiera de las dos formas. Por tanto, no hay una asociación categórica de cada zona con un lexema u otro, sino una preferencia por uno sin exclusión del otro. En este caso las diferencias son marcadas e indudablemente significativas. Pero con carácter general, y más aún en casos dudosos donde las diferencias no son tan claras, conviene acompañar observaciones de este tipo con pruebas estadísticas que muestren que las tendencias son significativas.

Lo que en el anterior ejemplo son diferencias geográficas en el léxico vale también para la variación diacrónica, social o de registro tanto en el léxico como en la gramática. En los estudios diacrónicos, por ejemplo, podemos buscar la primera o la última ocurrencia documentada de una forma lingüística, pero más interesante aún es trazar cómo aumenta o disminuye a lo largo del tiempo su frecuencia de uso. Para ello es importante, de nuevo, que comparemos frecuencias normalizadas y no frecuencias absolutas. Lamentablemente, ni CREA ni CORDE proporcionan frecuencias relativas y es difícil recuperar los datos para calcularlas uno mismo, aunque se han propuesto métodos alternativos de cálculo (Molina Salinas y Sierra Martínez 2015). En esta faceta concreta, los gráficos de CdE-hist, como el de la Figura 2, tienen la ventaja de que comparan frecuencias relativas (por millón) y no frecuencias absolutas.

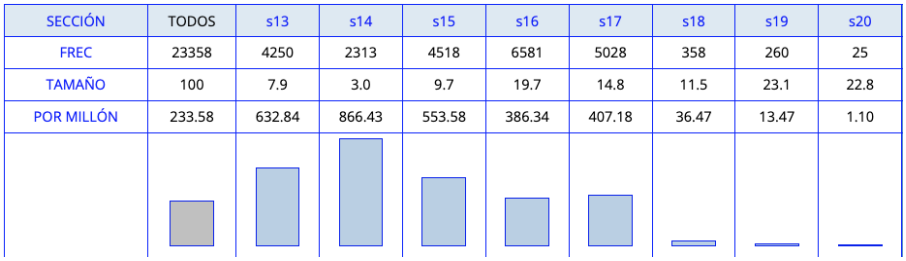

Figura 2. Frecuencias absolutas y relativas de *agora* en CdE-hist, por siglos

Las variaciones de frecuencia aparecen por todas partes en el sistema lingüístico y dan lugar a algunos principios generales interesantes. En toda lengua hay elementos que son muy frecuentes y elementos que son muy raros, de acuerdo con la ley de Zipf, la cual establece relación inversa entre la frecuencia de una palabra y su rango en el orden de frecuencias. Esto quiere decir que, si ordenamos las palabras de cualquier corpus de más a menos frecuente, en los primeros puestos tendremos palabras de uso muy frecuente y

ISSN: 1139-8736 https://infoling.org/elies/

rápidamente van bajando las frecuencias de cada unidad hasta el punto de que encontraremos solo un ejemplar de alrededor de la mitad de las unidades. Por otro lado, sea cual sea el corpus utilizado, en los primeros puestos de la lista encontraremos siempre palabras gramaticales de muy alta frecuencia, como observamos en la Tabla 3, que contiene las palabras más frecuentes de diferentes corpus del español.

	CORPES		**EsTenTen18**		**ARTHUS**		**ESLORA**	
1	de	63.009	de	59.688	de	50.449	y	39.307
2	la	39.785	la	33.503	la	36.323	de	34.228
3	que	31.472	que	26.244	que	34.077	no	31.988
4	el	29.502	el	24.865	y	28.141	que	28.726
5	y	27.901	en	24.695	el	25.081	a	27.753
6	en	27.551	y	24.560	a	24.727	el	23.756
7	a	21.456	a	17.400	en	22.816	la	21.868
8	los	16.367	los	13.911	no	17.884	en	18.158
9	se	12.990	del	10.468	los	12.718	es	15.818
10	un	11.411	se	10.194	se	12.426	sí	14.863

Tabla 3. Las diez palabras más frecuentes en cuatro corpus del español.
Frecuencias relativas (por millón)

Las diferencias que se observan en la Tabla 3. entre unos corpus y otros son muy poco significativas. De esos cuatro corpus, el que parece diferenciarse algo más de los otros es ESLORA, quizá por estar constituido por muestras de español oral mientras que los otros tres son mayoritariamente de español escrito. Lo cual nos indica una vez más que en la frecuencia de las formas en un corpus podemos encontrar algunas constantes (por ejemplo, la alta frecuencia de elementos gramaticales) junto con variabilidad debida a múltiples factores, tales como el medio oral o escrito, la procedencia geográfica, la variación diacrónica, etc.

Si observamos muestras relativamente pequeñas o textos individuales, lógicamente lo que más puede variar son los elementos léxicos, que están mucho más condicionados por la temática del texto. De los listados de frecuencia deriva un producto típico de la lingüística de corpus consistente en la obtención automática de las palabras clave de un texto o conjunto de textos, para lo que existen dos procedimientos básicos: o bien seleccionar las palabras más frecuentes de entre las de contenido léxico una vez eliminadas las palabras gramaticales, o bien seleccionar las que muestren una diferencia estadísticamente significativa respecto de la frecuencia relativa en un corpus más amplio que se toma como referencia. Para representar estos datos es práctica común el uso de "nubes de palabras", un tipo de gráfico que destaca las palabras clave de un texto, asignando mayor tamaño a las de mayor frecuencia como en la Figura 3:

Figura 3. Nube de palabras de un borrador de este mismo artículo

ISSN: 1139-8736 https://infoling.org/elies/

La ley de Zipf también tiene otras repercusiones prácticas para el análisis de corpus: en cualquier corpus obtendremos muchos ejemplares (a veces, más de los que somos capaces de analizar en detalle) de las unidades más frecuentes; pero apenas encontraremos ejemplares de las menos frecuentes. Si, por ejemplo, estamos haciendo un diccionario basado en un corpus cerrado, es probable que de muchas palabras apenas encontremos ejemplos en el corpus. Y, debido precisamente a la ley de Zipf, duplicar el tamaño del corpus no implica ni que dupliquemos el número de palabras diferentes registradas ni que se duplique el número de ejemplares de cada palabra registrada.

4.3. Frecuencias de coaparición: de colocaciones a perfiles combinatorios

Junto con las frecuencias de las unidades y su distribución en relación con diferentes parámetros de variación, la otra gran dimensión de análisis en lingüística de corpus son las relaciones de co-ocurrencia o coaparición, que tienen que ver con cómo se combinan unas unidades con otras en los textos. Sabemos que el lenguaje es un sistema combinatorio en el que los signos básicos, palabras o morfemas se unen para formar estructuras complejas: frases, oraciones, textos. Y sabemos, también, que no todas las combinaciones imaginables son posibles y que para separar lo que es posible en una lengua de lo que no lo es se elaboraron en lingüística los conocidos conceptos de gramaticalidad y de aceptabilidad. Pues bien, como ya hemos mencionado en el apartado anterior, la lingüística de corpus no tiene ni puede tener como objetivo determinar lo que es posible en una lengua, que es algo que no se puede extraer de ningún corpus, sino estudiar qué es lo que está documentado en un corpus y con qué frecuencia. Por tanto, en el caso de las coocurrencias, no se tratará de encontrar combinaciones posibles o imposibles sino de combinaciones más o menos probables.

En la observación de coapariciones lo que se revela es la dialéctica entre lo que Sinclair (1991) ha formulado como el principio de la elección abierta [“the open choice principle”] frente al principio de idiomaticidad [“the idiom principle”]. Por un lado, tenemos estructuras sintácticas como V + N en las que teóricamente podemos escoger cualquier nombre o cualquier verbo en las posiciones correspondientes. Por otro lado, los usuarios de la lengua tienen a su disposición un número grande de combinaciones ya construidas o semiconstruidas, que se repiten de unos textos a otros y aparecen frecuentemente en el corpus. Son fórmulas prefabricadas que no tienen por qué crearse *ex novo*, aunque su significado global pueda ser composicional, como en los sintagmas libres, o no serlo, como en las locuciones. En esta línea, la lingüística de corpus ha destacado que existen límites difusos entre las combinaciones fijas, las combinaciones preferentes más o menos restringidas y las combinaciones más abiertas con significado composicional.

Para estudiar las combinaciones de palabras, incluso en corpus que no tienen ningún tipo de anotación gramatical, se han desarrollado ciertas técnicas para resumir automáticamente la distribución de una unidad, entre las cuales destacan la búsqueda de n-gramas y el estudio de colocaciones. Los n-gramas (bigramas, trigramas, etc.) son cadenas de dos o más palabras consecutivas de uso frecuente. Es fácil generar automáticamente listas con todas las combinaciones de dos palabras, tres palabras, etc. y ordenarlas por su frecuencia. Las combinaciones que más se repiten son en buena medida fórmulas prefabricadas, aunque no necesariamente son locuciones, pues su significado global puede ser composicional o no serlo. Por ejemplo, en un corpus de recetas de cocina de elaboración propia encontramos que los trigramas más frecuentes son, por este orden,

ISSN: 1139-8736 https://infoling.org/elies/

aceite de oliva, un poco de y *salsa de tomate.* Esto nos proporciona información en parte extralingüística sobre qué ingredientes solemos utilizar para cocinar, pero también lingüística sobre expresiones que aprendemos y usamos como combinaciones prefabricadas en determinados contextos discursivos.

En lingüística de corpus, se llama *colocación* a la relación entre palabras que aparecen juntas con mayor frecuencia de lo esperable por pura casualidad. En este caso no se trata solo de palabras que aparecen una al lado de otra, como en los n-gramas, sino que se busca en las proximidades, entendiendo por proximidad una "ventana" de *n* palabras a izquierda o derecha de la palabra buscada de tamaño variable y que suele fijarse en un máximo de 5 palabras a la izquierda o a la derecha. Y en ese margen lo que se buscan no son simplemente coapariciones frecuentes, sino que la coaparición sea estadísticamente significativa, comparando la frecuencia observada con la teóricamente esperable a partir de la frecuencia total que tienen en el corpus las palabras combinadas. Por ejemplo, si buscamos en CORPES sin ningún tipo de restricción cuáles son las palabras que aparecen más frecuentemente en las proximidades de un verbo como *vencer* encontraremos que las más frecuentes son el artículo *el* y las preposiciones *a* y *de*, que son las palabras más frecuentes en todo el corpus, y por tanto no puede decirse que hayan sido atraídas en particular por ese verbo. En cambio, una palabra como *timidez* tiene más probabilidades de aparecer en el entorno de *vencer* que en otras partes del corpus, por lo que podemos decir que *vencer* la atrae y que *timidez* es un colocado de este verbo.

Existen diferentes medidas estadísticas de asociación (Church y Hanks 1990) que comparan de distintas maneras la frecuencia observada en una colocación con la frecuencia esperada. El buscador de CORPES ofrece varias de estas medidas para elegir cómo ordenar los resultados. En la Tabla 4 se han extraído manualmente los nombres abstractos que coaparecen más con el verbo *vencer* y se han ordenado por MI ('mutual information'), que utiliza una fórmula que divide frecuencia observada y esperada. Otras fórmulas de cálculo (t-score, LL-simple) situarían en primer lugar la palabra *resistencia*, otorgando más peso a las combinaciones con mayor frecuencia absoluta.

	Frec	**MI**	**LL simple**	**T score**
timidez	84	9,3	399,37	9,16
repugnancia	23	8,85	102,97	4,79
reticencia	38	8,83	169,76	6,16
obstáculo	226	8,5	964,74	15,03
adversidad	38	8,35	158,62	6,16
pereza	34	8,2	138,85	5,83
escollo	24	8,2	98,06	4,89
cansancio	120	8,06	480,04	10,95
resistencia	313	7,94	1230,43	17,63
dificultad	163	6,28	478,38	12,61

Tabla 4. Nombres abstractos que aparecen en las proximidades del verbo VENCER en el *Corpus del español del siglo XXI (CORPES).* Elaboración propia

En la exploración de CORPES para buscar las coapariciones del verbo *vencer* de la Tabla 4, antes de la extracción manual de los nombres abstractos, el primer criterio (MI) situaba en los primeros puestos de la lista una serie de resultados numéricos (tipo "2-1") y una serie de nombres de equipos de la NBA (como "Mavericks"), que aparecen muy raramente en ese corpus, pero de manera frecuente en las proximidades del verbo *vencer*.

ISSN: 1139-8736 https://infoling.org/elies/

Los otros criterios de ordenación de CORPES sitúan en las primeras posiciones palabras gramaticales muy frecuentes. La moraleja del ejemplo es que si queremos obtener resultados significativos sobre colocaciones en un corpus no quedará más remedio que evaluar los resultados que proporciona cada fórmula de cálculo. De las diferentes fórmulas alternativas, el cálculo de "información mutua" (MI) suele ofrecer buenos resultados si se descartan los casos de menor frecuencia.

La tarea de encontrar colocaciones significativas se verá facilitada enormemente si contamos con un buen sistema de exploración de corpus que nos permita ajustar los parámetros de búsqueda. Por ejemplo, el sistema de búsqueda del *Corpus del español*, además de fijar unos valores mínimos de frecuencia y de relevancia, permite entre otros ajustes seleccionar la ventana de búsqueda y la categoría del colocado, de modo que podemos buscar nombres que aparecen en hasta tres posiciones a la derecha de *vencer*, con lo que previsiblemente obtendremos objetos directos típicos de ese verbo: las primeras posiciones son ocupadas por *repugnancia*, *obstáculo*, *resistencia* y *dificultad*; lo que se corresponde bastante bien con nuestras intuiciones sobre la combinatoria de este verbo. Con todo, esto no nos libra del proceso de evaluación. En las primeras posiciones de la lista aparecían también otros sustantivos como *batalla*, *lucha* o *plazo*, que no solo pertenecen a otros campos semánticos, sino que mantienen diferente relación sintáctica con el verbo: complemento de lugar (*vencer en la batalla/lucha*) o sujeto (*vencer el plazo*). Sería deseable que los procedimientos de exploración de colocaciones fueran sensibles a la estructura sintáctica y semántica (por ejemplo, nombres abstractos que ocurren como objeto directo de un verbo dado), pero esto está supeditado a la existencia de corpus anotados sintácticamente.

Los corpus y sistemas de consulta más difundidos (incluyendo el procedimiento de elaboración de "*word sketches*" por parte de la plataforma *Sketch Engine*, más sofisticado de lo habitual pero basado en una gramática algo rudimentaria) nos ofrecen un sucedáneo basado en la categoría morfosintáctica de las palabras a la derecha y a la izquierda de la palabra nodo. Aunque presenta limitaciones, podemos obtener igualmente una muy útil visión general de la distribución de una palabra dada. La Figura 4 nos muestra parte de un *word sketch* del verbo *vencer* con las palabras que aparecen típicamente junto a él como modificadores (adverbios), objetos (nombres a la derecha del verbo o nombres modificados por el participio pasivo) u elementos coordinados.

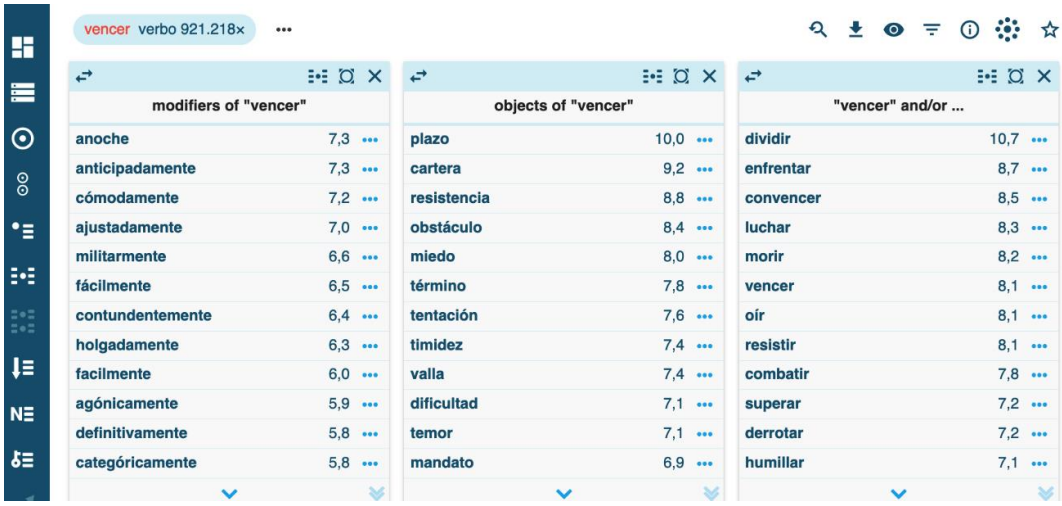

Figura 4. *Word Sketch* del verbo VENCER en el corpus esTenTen18 (Sketch Engine)

ISSN: 1139-8736 https://infoling.org/elies/

A este respecto, cabe señalar que el concepto de colocación utilizado en lingüística de corpus, basado en la frecuencia observada en los textos, ha sido criticado con el argumento de que combinaciones frecuentes como la de cierto verbo con cierto sustantivo (por ejemplo, *leer* + *libro*) no nos proporcionan propiedades lingüísticas de esas palabras sino propiedades externas de las entidades del mundo a las que se aplican ciertas acciones. Esa es la tesis de Bosque (2001), quien propone usar el término colocación para referirse solo a combinaciones léxicas restringidas idiomáticamente, como por ejemplo *vino tinto* (y no *vino rojo*), pero no para referirse a cualquier combinación frecuente. Sin embargo, las colocaciones, entendidas como en lingüística de corpus como coapariciones más frecuentes de lo esperable por el azar, son imprescindibles para conocer el significado y uso de las palabras. Aportan conocimiento idiomático, normas de uso específicas de una lengua particular, como, por ejemplo, que el *vino* es *blanco* o *tinto*, no amarillo y rojo. Así mismo, aportan conocimiento social-discursivo: qué cosas se suelen decir de qué; por ejemplo, qué adjetivos y nombres valorativos se usan en la prensa alrededor de la palabra 'musulmán' y, por tanto, qué prejuicios y actitudes se transmiten hacia los musulmanes en el discurso publicado (Baker et al. 2013). Y, ciertamente, aportan también conocimiento de la realidad designada por las palabras, como que *leer* se combina frecuentemente con *libro*; algo que por supuesto puede cambiar si cambia la realidad: así, en CdE-hist los nombres que se colocan más frecuentemente a la derecha del verbo *leer* son *libro, carta, periódico* y *novela*; pero en el más reciente CdE-web la misma búsqueda da como resultado *libro, artículo, comentario* y *blog*. En definitiva, por mucho que estén determinadas por la realidad designada, son esas coapariciones las que utilizamos para interpretar y aprender el significado de las palabras.

Como en el corpus no hay significado como tal, solo hay cadenas de caracteres y frecuencias de aparición y coaparición, son esas frecuencias las que nos están informando sobre cuál es el significado de las palabras y en qué se parecen y en qué se diferencian. La lingüística de corpus se adhiere así a la hipótesis distribucional (elaborada por Bloomfield y seguidores, en particular por Z. S. Harris) según la cual las entidades lingüísticas que tienen una distribución similar tienen significados similares y, viceversa, las diferencias de significado se reflejan en diferencias de distribución. De esta manera, podemos hacer con el análisis de corpus una semántica basada en el uso utilizando métodos estadísticos. En el procesamiento del lenguaje (PLN) o lingüística computacional, este principio permite desarrollar métodos automáticos a partir de corpus (anotados o no) para agrupar palabras semánticamente similares porque tienen distribuciones similares, infiriendo de ello semejanzas y diferencias de significado (Martí Antonín 2018).

De ese principio general, que las diferencias y semejanzas de significado se manifiestan en diferencias y semejanzas de distribución, deriva también el concepto de *perfil combinatorio*, que da lugar a un método derivado usado en semántica lingüística basada en corpus consistente en seleccionar muestras de la palabra o palabras analizadas y anotar manualmente rasgos formales y funcionales que presentan los ejemplos contextualizados de la muestra y luego aplicar métodos estadísticos multifactoriales de agrupación y discriminación que permitan describir la estructura de un campo semántico, o la de una palabra polisémica o mostrar las diferencias entre cuasi-sinónimos a partir de su perfil combinatorio ['*Behavioral Profile*'] (Gries y Divjak 2006, Gries 2010a, Glynn y Robinson 2014). La Figura 5 muestra en una estructura arbórea las semejanzas relativas entre verbos en ruso del campo semántico "intentar" de acuerdo con este método.

ISSN: 1139-8736 https://infoling.org/elies/

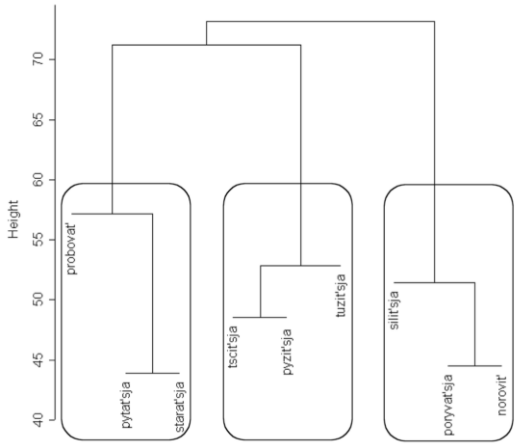

Figura 5. Dendrograma de verbos intencionales en ruso según su perfil combinatorio (Divjak y Gries 2006: 38)

4.4. Patrones sintagmáticos

En lo que concierne a la relación de la lingüística de corpus con los estudios sintácticos, como se ha comentado un poco más arriba, muchos corpus contienen anotación morfosintáctica de las palabras, pero es raro que los grandes corpus contengan una anotación sintáctica detallada, algo reservado a corpus relativamente pequeños con anotación total o parcialmente manual. La anotación sintáctica de corpus, como por ejemplo la integrada en ADESSE, permite investigar en detalle fenómenos sintácticos como la marcación variable del objeto (García-Miguel 2015), esto es, la presencia variable de la preposición *a*, la duplicación pronominal o la elección de caso acusativo (*lo*) o dativo (*le*) con objetos directos e indirectos. Fenómenos de este tipo pueden investigarse también en un corpus mínimamente anotado, pues con búsquedas relativamente simples podemos documentar el uso y distribución de cadenas de caracteres como *le* y *lo.*

En corpus con anotación morfosintáctica podemos explorar, como se ha indicado antes, no solo palabras y frases particulares sino también categorías gramaticales en sí mismas –por ejemplo, cuál es la frecuencia relativa de nombres y verbos en un (sub)corpus– o en combinación con otros elementos –por ejemplo, qué nombres aparecen a la derecha del verbo *vencer* o cuál es la distribución de la construcción Adverbio + Posesivo del tipo *detrás mío* (Eddington 2017). Ciertamente, la metodología basada en corpus no suele centrarse en reglas combinatorias muy generales, sino en la exploración de construcciones particulares de un nivel de abstracción relativamente bajo.

Algunas aproximaciones teóricas a la gramática se basan en patrones sintagmáticos que emergen del corpus. Más que en reglas o principios combinatorios abstractos de aplicación general, tales patrones combinan elementos abstractos o esquemáticos con elementos léxicos específicos o clases de elementos léxicos de manera que se difumina la distinción entre léxico y gramática. En esta línea podemos mencionar la *Pattern Grammar* de Hunston y Francis (1998, 2000), el '*Corpus Pattern Analysis*' de P. Hanks (Hanks y Pustejovski 2005) y otros trabajos que destacan que la descripción de la combinatoria de una unidad léxica debe incluir junto con la valencia sintáctico-semántica también información colocacional (Butler 2001). La técnica de análisis consiste en observar instancias de uso en corpus y agrupar las líneas de concordancia según patrones

ISSN: 1139-8736 https://infoling.org/elies/

sintagmáticos semánticamente motivados, de modo que los significados se asocian con contextos prototípicos. En el caso de los verbos, los patrones no incluyen simplemente la estructura valencial sino también detalles contextuales potencialmente relevantes para el significado, como puede ser la presencia de ciertos nombres o clases de nombres, de cierta preposición o incluso la presencia o ausencia de determinante junto al nombre complemento (por ej. *tener lugar* es distinto de *tener un lugar*). En (4) tenemos un ejemplo en inglés de Hunston y Francis con los patrones que son parte de la conducta típica del sentido relevante de cada verbo. En (5) tenemos parte de la entrada del verbo *preocupar* en Verbario, siguiendo el modelo de *Corpus Pattern Analysis* utilizado en el PDEV por P. Hanks.

(4) *we tended to think of ourselves as not having to worry about those freaks down there* [‘tendíamos a pensar en nosotros mismos como si no tuviéramos que preocuparnos por esos monstruos de abajo’]
Verb patterns:
tended V to-inf
think V *of* n *as* n/-ing
worry V *about* n (Hunston & Francis 1998: 64-65)

(5) Patrones de PREOCUPAR en *Verbario*
1a [[Humano | Eventualidad]] *preocupar* [[a Humano]]
1b [[Humano 1]] *preocuparse* (*por/de* Humano 2 | Eventualidad)

El análisis llamado colostruccional, propuesto por Stefanowitsch y Gries (2003), combina el concepto de colocación con el concepto de construcción. El objetivo es observar en una construcción dada qué lexemas particulares entran preferentemente en los huecos abiertos por la construcción, fuertemente atraídos por ella. El análisis puede aplicarse tanto a construcciones muy específicas, como [N *waiting to happen*] (lit. ‘N esperando a suceder’), para las que podemos explorar en un corpus del inglés qué nombres pueden rellenar el hueco que queda libre, como a esquemas puramente abstractos como la construcción ditransitiva ([S_{agent} V O_{rec} O_{th}] en inglés, [Suj – V – OD – OI] en español), para la que podemos explorar en corpus cuáles son los verbos atraídos por ese esquema. En este último caso se comprueba que no cabe cualquier nombre ni cualquier verbo en ese esquema, sino que la construcción está claramente asociada con el verbo *dar* y otros verbos semánticamente relacionados con la transferencia, tanto en inglés como en español (Tabla 5). Stefanowitsch y Gries utilizan una prueba estadística, el test exacto de Fisher, para calcular la fuerza de la asociación [*collostructional strength*], aunque en este caso concreto la observación de frecuencias absolutas puede proporcionarnos resultados similares.

Collexeme	**Coll. strength**
give (461)	0
tell (128)	1.6E-127
send (64)	7.26E-68
offer (43)	3.31E-49
show (49)	2.23E-33
cost (20)	1.12E-22
teach (15)	4.32E-16
award (7)	1.36E-11

VERBO	**Frec**
dar	1184
decir	598
hacer	486
contar	308
pedir	272
preguntar	220
poner	144
permitir	124

ISSN: 1139-8736 https://infoling.org/elies/

Tabla 5. Verbos asociados con la construcción ditransitiva en inglés (Stefanowitsch y Gries 2003: 229) y en español (ADESSE. Elaboración propia)

La moraleja es que los esquemas sintácticos abstractos se asocian con elementos léxicos particulares. Y viceversa, los elementos léxicos no tienen solo una manera de construirse. Los verbos no tienen una valencia fija, sino que admiten diferentes esquemas construccionales y tienen una probabilidad mayor o menor de aparecer en un esquema o en otro (García-Miguel 2005), como podemos observar en el ejemplo de la Figura 6 con las frecuencias de los esquemas valenciales del verbo *vencer* en ADESSE.

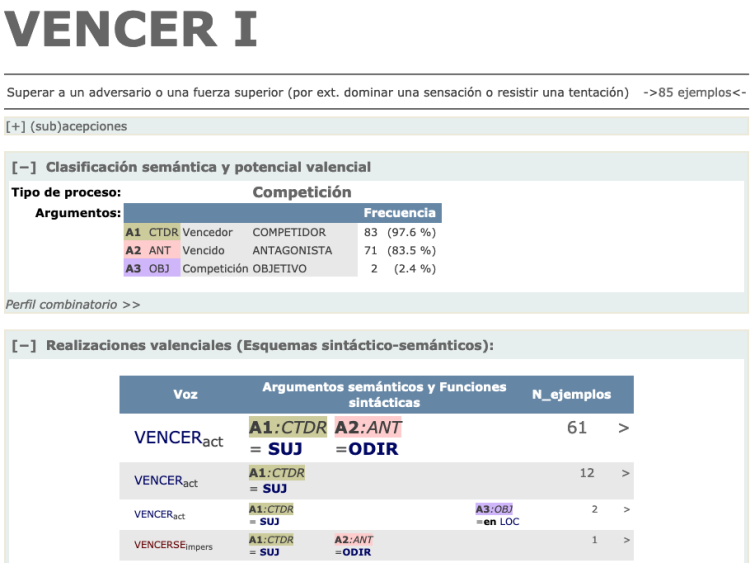

Figura 6. Esquemas sintáctico-semánticos de *VENCER* (ADESSE)

En definitiva, lo que más y mejor solemos hacer con un corpus es buscar patrones de coocurrencia, de elementos léxicos con elementos léxicos (colocaciones), de elementos y esquemas gramaticales con elementos léxicos (por ej. colostrucciones), de elementos léxicos con esquemas gramaticales (valencia), etc. Podemos generalizar el concepto de *perfil combinatorio* para referirnos al conjunto de probabilidades de coocurrencia de una unidad lingüística cualquiera con elementos léxicos y gramaticales de cualquier nivel de generalidad. Estas probabilidades varían según el contexto, por lo que deberíamos añadir factores sociales y discursivos y aplicar técnicas estadísticas de análisis multifactorial, algo que no desarrollaremos por los límites de espacio de este artículo.

4.5. Final de sección

En resumen, virtualmente cualquier aspecto del léxico y de la gramática de las lenguas puede estudiarse en un corpus. Pero en la práctica es lógico que existan ciertos sesgos motivados por la relativa facilidad para digitalizar textos, anotarlos y explorarlos con ayuda de la informática. Una tendencia general es la de focalizar el interés en la frecuencia de unidades y patrones que son directamente observables en el corpus con ayuda de ordenador. Pero tales observaciones necesitan acompañarse por parte del lingüista de procesos de abstracción tanto en la clasificación como en la interpretación de los resultados obtenidos. Según Barlow (2011: 7-8) las principales aportaciones de la lingüística de corpus a la investigación lingüística han sido tres: (i) destacar la

ISSN: 1139-8736 https://infoling.org/elies/

importancia de las co-apariciones, (ii) focalizarse en la frecuencia y, por tanto, en formas típicas de expresión más que en formas posibles, y (iii) cuantificar el rango de variación en las lenguas. Por otro lado, los estudios guiados por corpus suelen presentar ciertos rasgos comunes (adaptado de Barlow 2011: 11):

- prominencia de las unidades léxicas en las descripciones;
- se recurre a categorías más abstractas o esquemáticas (categorías, relaciones) solo en caso necesario;
- las semejanzas y diferencias de significado se asocian con la distribución;
- los procesos de interpretación y la conexión con aspectos cognitivos, sociales, pragmáticos o discursivos no siempre se establecen de manera sistemática.

5. Lingüística de corpus y teoría(s) lingüística(s)

La lingüística de corpus es ante todo una metodología cuyos conceptos y resultados son compatibles con diferentes modelos lingüísticos, de modo que cualquier modelo teórico puede utilizar, en mayor o menor medida, datos de corpus. Sin embargo, es cierto que algunas perspectivas se han mostrado más proclives que otras a utilizar datos de corpus y que los datos de corpus pueden utilizarse de diferentes maneras con distinta repercusión en la teoría. Puede distinguirse (Tummers et al. 2005: 234-236) entre lingüística *ilustrada con corpus* y lingüística *basada en corpus*. La primera toma ejemplos de corpus para complementar datos y realiza análisis basados más bien en la intuición. En la segunda, la evidencia empírica y las tendencias encontradas en el uso constituyen el núcleo del análisis y definen el modelo resultante, y en estas aproximaciones basadas en corpus las técnicas estadísticas constituyen una parte esencial del análisis.

Muchos lingüistas que hacen sintaxis más o menos formal piensan que con la intuición es suficiente para elaborar modelos teóricos de lo que es posible en una lengua, o en general en las lenguas. Entre ellos se encuentran lingüistas tan prominentes como Chomsky o Tesnière. Este último afirma explícitamente que la introspección debe convertirse en una de las piezas maestras de la investigación sintáctica (Tesnière 1959: 37). Pero también hay quienes hacen sintaxis más o menos formal y han incorporado nociones de gradación, frecuencia y probabilidad basadas en la observación del uso real, por tanto, en corpus (por ejemplo, Manning 2003, Aarts 2007, Bresnan y Hay 2008).

Con todo, con lo que parece mostrar mayor compatibilidad la lingüística de corpus es con los modelos teóricos que se dicen "basados en el uso" (Barlow y Kemmer 2000), esto es, los modelos cognitivos y funcionales. Si se pretende basar el modelo en el uso hay que observar el uso real y el método que surge como primer candidato es, obviamente, la lingüística de corpus. Destacaremos en particular la conexión de la lingüística de corpus con el contextualismo británico (incluyendo la Lingüística Funcional de Halliday) y con la Lingüística Cognitiva.

Donde más se desarrolló inicialmente la lingüística de corpus fue en Gran Bretaña, en buena parte como continuación de la perspectiva contextualista de Firth (1890-1960) a quien debemos el aforismo "a una palabra se la conoce por sus acompañantes" ["You shall know a word by the company it keeps"] (Firth 1957: 179), que está en la base de su concepto de colocación. Firth influye en la llamada lingüística de corpus neo-firthiana

ISSN: 1139-8736 https://infoling.org/elies/

(Sinclair, Hunston, Teubert, …), con su aproximación inductiva "guiada por corpus" [*corpus-driven*]; pero también influye en otros lingüistas británicos con una orientación más "basada en corpus" [*corpus-based*] (Aarts, Biber, Leech, McEnery). Esta distinción entre dos aproximaciones aparentemente contrapuestas fue formulada por Tognini-Bonelli (2001), según la cual, en las aproximaciones "basadas en corpus", la teoría y las categorías analíticas se establecen previamente y se comprueban o ejemplifican en el corpus; mientras que, en las aproximaciones "guiadas por corpus", la teoría y las categorías analíticas derivan de los datos, emanan del corpus, siguiendo un método puramente inductivo. Esta última perspectiva es la defendida por la propia Tognini-Bonelli y por figuras destacadas como John Sinclair (Sinclair 2004). Por el contrario, otros especialistas (McEnery et al. 2006: 8-10) aducen que la distinción entre esas dos aproximaciones se ha exagerado y que es más bien borrosa; lo que nos da pie a propugnar una metodología que recorra continuamente el camino de ida y vuelta entre teoría y datos.

La insistencia de algunos lingüistas británicos, especialmente de Sinclair, en seguir un método estrictamente inductivo que evite cualquier categoría que no emane del propio corpus les da a estas aproximaciones una apariencia ateórica o carente de un modelo teórico elaborado. Por el contrario, no tiene nada de ateórica la Lingüística Sistémico-Funcional de Michael Halliday, también notablemente influida por Firth. La teoría de Halliday destaca por su capacidad para integrar la explicación del sistema lingüístico en relación con el texto y el contexto. Aunque esta teoría no está necesariamente ligada al uso de corpus (pero sí se usan textos reales como ilustración de la teoría, más que oraciones inventadas), Halliday siempre defendió la relevancia teórica de la lingüística de corpus y la importancia de la frecuencia en la conformación del sistema lingüístico. Para él, el significado deriva de una red de sistemas de elecciones (de ahí el término "sistémica") y una parte importante del significado de un término cualquiera es la probabilidad de escogerlo frente a otro(s). Así, por ejemplo, el significado de 'negativo' no es simplemente 'no positivo' sino 'no positivo con una razón de probabilidades de uno a nueve'. Como las probabilidades varían según el contexto, las elecciones gramaticales pueden significar cosas diferentes en registros diferentes (Halliday 1991: 32-33).

La influencia de Firth se refleja también en el interés creciente por los estudios basados en corpus de las relaciones entre lenguaje, contexto, sociedad y cultura (Gabrielatos 2021).

Un campo donde la lingüística de corpus ha encontrado terreno muy abonado es el de la Lingüística Cognitiva (Tummers et al. 2005; Gries 2010b: 334-336). Frente a la dicotomía tradicional entre sistema (lengua, competencia) y uso (habla, actuación), la Lingüística Cognitiva propone que el sistema está conformado desde el principio por el uso, y que las estructuras lingüísticas emergen del uso individual y colectivo del lenguaje. Y la lingüística basada en el uso necesita cuantificación y análisis estadístico (Tummers et al. 2005: 234). En este sentido, un concepto clave para el modelo de gramaticalidad basado en el uso es el de *entrenchment* ['consolidación mental']: las estructuras lingüísticas nuevas se consolidan progresivamente gracias al uso repetido y las unidades estarán más o menos consolidadas dependiendo de su frecuencia de ocurrencia (Langacker 1987: 59). Este concepto de consolidación mental reformula el concepto de gramaticalidad (que además deja de ser una cuestión de sí o no, para convertirse en algo gradual) y le proporciona una base empírica que puede contrastarse con las frecuencias observadas en un corpus. Sin embargo, hay que cuidarse de establecer una relación directa entre frecuencia simple y consolidación mental (Arppe et al. 2010: 8-10). Otras facetas

ISSN: 1139-8736 https://infoling.org/elies/

en las que encaja bien la lingüística de corpus con la Lingüística Cognitiva están en el estudio de la categorización basada en ejemplares y en los efectos de prototipicidad en la polisemia (Mukherjee 2004; Glynn y Robinson 2014), nuevamente con la advertencia de que no se puede equiparar prototipicidad y frecuencia simple. Y la concepción del lenguaje como un sistema complejo que no se puede entender haciendo abstracción de factores contextuales y sociales (algo también esencial en la perspectiva cognitivista) se plasma en estudios cuantitativos basados en corpus que aplican modelos multifactoriales que incluyen explícitamente la variación social y cultural (Heylen et al. 2008).

Tanto la Gramática Cognitiva como su pariente próximo la Gramática de Construcciones (Goldberg 2003) entienden la gramática como una red de unidades convencionales que asocian formas y significados a diferentes niveles de esquematicidad y fijación. En una línea similar, perfectamente compatible con la de la Lingüística Cognitiva, tenemos trabajos como los de Bybee, Hopper y otros (Bybee y Hopper 2001; Bybee 2007, etc.), con la idea de que la gramática emerge del uso, es desde su raíz variable y probabilística y se constituye a partir de instancias específicas de uso mediante procesos de rutinización y esquematización basados en la categorización de ejemplares. En esta aproximación se destaca especialmente el papel de la frecuencia en los mecanismos de cambio y gramaticalización (Bybee 2003).

6. (A modo de) conclusión

Cualquier lingüista, casi con independencia de su campo de investigación, puede explorar corpus de distintas fuentes con la seguridad de que, en poco tiempo, podrá acumular una gran cantidad de datos relevantes (muchas veces mezclados con datos irrelevantes) y que entre ellos habrá muchos datos sorprendentes. Pero no se debe olvidar que el objetivo de cualquier investigación es avanzar en el conocimiento del objeto de estudio, no simplemente acumular datos. Y en ese objetivo deben evaluarse siempre las limitaciones que pueda tener tanto el corpus utilizado como la propia metodología basada en corpus, por lo que idealmente debería complementarse con otros métodos como los que se exponen en otros capítulos de este volumen. Como indicamos más arriba, se ha comparado la importancia para la lingüística de la informatización de datos textuales con lo que supuso el telescopio para la astronomía; pero cada herramienta tiene su función y su utilidad dependerá de lo que queramos hacer: tan ridículo como criticar un telescopio por no servir de microscopio sería el criticar la lingüística de corpus por no hacer aquello para lo que no está concebida (McEnery et al. 2006: 121).

Para conseguir adecuación descriptiva necesitamos corpus textuales representativos de la lengua o variedad de lengua objeto de estudio. También necesitamos anotaciones detalladas, pues no siempre es suficiente con los resultados que nos proporciona la búsqueda de secuencias de caracteres de palabras o de categorías morfosintácticas. Las categorías usadas en la anotación de corpus deben estar fundamentadas teóricamente y deben revisarse a la luz de la evidencia proporcionada por el propio corpus. Debemos movernos continuamente de las categorías analíticas al corpus y del corpus a las categorías analíticas. En mi opinión, se observa en las últimas décadas cierta aproximación entre los estudios más teóricos y los más empiristas, aunque probablemente aún queden muchos pasos por dar hasta llegar a la integración total. Los estudios de corpus incluyen propiedades sintácticas y semánticas cada vez más refinadas y los

ISSN: 1139-8736 https://infoling.org/elies/

estudios más teóricos están prestando progresivamente más atención a los datos empíricos y a las variaciones que se observan en el uso. Entre las principales aportaciones de la lingüística de corpus a la lingüística contemporánea está, además de la puesta a disposición de una ingente cantidad de datos, el desarrollo de métodos cuantitativos que ponen en relación la frecuencia de uso con múltiples factores co-textuales y contextuales.

7. Referencias de corpus y otros recursos lingüísticos mencionados

ADESSE: Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español. http://adesse.uvigo.es
Ameresco: Corpus América y España Español Coloquial. http://esvaratenuacion.es/
AnCora: Multilingual and multilevel ANnotated CORporA. http://clic.ub.edu/corpus/
AntConc: [software] https://www.laurenceanthony.net/software/antconc/
Auslan Corpus: Australian Sign Language Corpus. https://www.auslan.org.au/about/corpus/
BDS: Base de Datos Sintácticos del Español Actual. http://www.bds.usc.es/
BNC: British National Corpus. http://www.natcorp.ox.ac.uk/
Brown corpus: The Standard Corpus of Present-Day Edited American English. https://varieng.helsinki.fi/CoRD/corpora/BROWN/
BSL Corpus: British Sign Language Corpus Project. https://bslcorpusproject.org/
CAES: Corpus de Aprendices de Español como lengua extranjera. http://galvan.usc.es/caes/
CdE-hist: Corpus del Español: Género/histórico. https://www.corpusdelespanol.org/hist-gen/
CdE-NOW: Corpus del Español: NOW (News on the web). https://www.corpusdelespanol.org/now/
CdE-web: Corpus del Español: web/dialectos. https://www.corpusdelespanol.org/web-dial/
CHILDES: Child Language Data Exchange System. http://childes.talkbank.org/
CLUVI Parallel Corpus. http://sli.uvigo.gal/CLUVI/
COLA: Corpus Oral del Lenguaje Adolescente. https://blogg.hiof.no/colam-esp/el-corpus-cola/
CORDE: Corpus Diacrónico del Español. http://corpus.rae.es/cordenet.html
CORELE: Corpus Oral de Español como Lengua Extranjera. http://www.lllf.uam.es/ESP/CORELE.html
CORLEC: Corpus Oral de Referencia de la Lengua Española Contemporánea. http://www.lllf.uam.es/ESP/Corlec.html
CorLSE: Corpus de la Lengua de Signos Española. https://corpuslse.es/
CORPES: Corpus del Español del siglo XXI. http://web.frl.es/CORPES/
Corpus del Español. http://corpusdelespanol.org/
Corpus Tècnic de l'IULA. https://www.upf.edu/web/iula/corpus-eines
Corpus Técnico do Galego. http://sli.uvigo.es/CTG/
CREA: Corpus de Referencia del Español Actual. http://corpus.rae.es/creanet.html
DGS-Korpus: Deutsche Gebärdensprache Korpus. https://www.sign-lang.uni-hamburg.de/dgs-korpus/
ELAN: [software] https://archive.mpi.nl/tla/elan

ISSN: 1139-8736 https://infoling.org/elies/

enTenTen20: Corpus of the English web.
https://www.sketchengine.eu/ententen-english-corpus/
ESLORA: Corpus para el estudio del español oral de Galicia. http://eslora.usc.es/
esTenTen18: Spanish web corpus 2018.
https://www.sketchengine.eu/estenten-spanish-corpus/
FreeLing: [software] http://nlp.lsi.upc.edu/freeling/
IULA Spanish Treebank. http://www.iula.upf.edu/recurs01_tbk_uk.htm
NGT Corpus: Corpus Nederlandse Gebarentaal. https://www.corpusngt.nl/
PDEV: Pattern Dictionary of English Verbs. https://pdev.sketchengine.eu/
PRESEEA: Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. https://preseea.linguas.net/Corpus.aspx
SenSem corpus: Sentence Semantics, Base de datos de semántica oracional.
http://grial.edu.es/sensem/corpus
Sketch Engine. http://www.sketchengine.eu
TEI: Text Encoding Initiative. http://www.tei-c.org/
TenTen corpus family. https://www.sketchengine.eu/documentation/tenten-corpora/
TreeTagger: [software] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
Val.Es.Co: Corpus Valencia, Español Coloquial
https://www.uv.es/corpusvalesco/consulta.html
Verbario: Semántica de los verbos en contexto. http://www.verbario.com

8. Referencias

Aarts, Bas. 2007. *Syntactic gradience: The nature of grammatical indeterminacy*. Oxford: Oxford University Press.

Arppe, Antti; Gilquin, Gaëtanelle; Glynn, Dylan; Hilpert, Martin; Zeschel, Arne. 2010. Cognitive Corpus Linguistics: five points of debate on current theory and methodology. *Corpora* 5.1: 1–27.

Baker, Paul; Gabrielatos, Costas; McEnery, Tony. 2013. Sketching Muslims: A corpus driven analysis of representations around the word "Muslim" in the British press 1998-2009. *Applied Linguistics* 34.3: 255–278.

Barlow, Michael. 2011. Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics* 16: 3–44.

Barlow, Michael; Kemmer, Suzanne, eds. 2000. *Usage-based models of language*. Stanford: CSLI.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8.4: 243–257.

Bresnan, Joan; Hay, Jennifer. 2008. Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua* 118.2: 245–259.

Bosque, Ignacio. 2001. Sobre el concepto de "colocación" y sus límites. *Lingüística Española Actual* 23.1: 9–40.

Butler, Christopher. S. 2001. A matter of GIVE and TAKE: corpus linguistics and the predicate frame. *Revista Canaria de Estudios Ingleses* 42: 55–78.

Bybee, Joan. 2003. Mechanism of change in grammaticization: the role of frequency. En B. Joseph y R. Janda, eds. *The Handbook of Historical Linguistics*. Oxford: Blackwell, pp. 602–623.

ISSN: 1139-8736 https://infoling.org/elies/

Bybee, Joan. 2007. *Frequency of use and the organization of language*. New York: Oxford University Press.
Bybee, Joan; Hopper, Paul, eds. 2001. *Frequency and the emergence of Linguistic structure*. Amsterdam: John Benjamins.
Church, Kenneth W.; Hanks, Patrick. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16.1: 22–29.
Davies, Mark. 2005. The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics* 10.3: 307–334.
De Benito Moreno, Carlota. 2019. Los corpus del español desde la perspectiva del usuario lingüista. *Scriptum digital* 8: 1–21.
Divjak, Dagmar; Gries, Stefan. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2.1: 23–60.
Dyson, Freeman J. 1997. *Imagined worlds*. Harvard University Press.
Eddington, David. 2017. Nominalized adverbs in Spanish: The intriguing case of *detrás mío* and its cohorts. *Research in Corpus Linguistics* 5: 47–55.
Fillmore, Charles J. 1992. "Corpus linguistics" or "Computer-aided armchair linguistics". En J. Svartvik, ed. *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, pp. 35–60.
Firth, John. 1957. *Papers in Linguistics.* Oxford University Press.
Gabrielatos, Costas. 2021. Bibliography of discourse-oriented corpus studies. http://ehu.ac.uk/docsbiblio.
García-Miguel, José M. 2005. Aproximación empírica a la interacción de verbos y esquemas construccionales, ejemplificada con los verbos de percepción. *Estudios de Lingüística* 19: 169–191.
García-Miguel, José M. 2015. Variable coding and object alignment in Spanish: A corpus-based approach. *Folia Linguistica* 49.1: 205–256.
García-Miguel, José M.; Vaamonde, Gael; González Domínguez, Fita. 2010. ADESSE, a Database with Syntactic and Semantic Annotation of a Corpus of Spanish. En *LREC2010 - Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta (Malta): ELRA, pp. 1903–1910.
Gilquin, Gaëtanelle; Gries, Stefan. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5.1: 1–26.
Glynn, Dylan; Robinson, Justyna A., eds. 2014. *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*. Amsterdam: John Benjamins.
Goldberg, Adele. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7.5: 219–224.
Gries, Stefan. 2009. *Quantitative Corpus Linguistics with R: A practical introduction*. Londres: Routledge.
Gries, Stefan. 2010a. Behavioral profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon* 5.3: 323–346.
Gries, Stefan. 2010b. Corpus linguistics and theoretical linguistics: A love–hate relationship? Not necessarily… *International Journal of Corpus Linguistics* 15.3: 327–343.
Halliday, M.A.K. 1991. Corpus studies and probabilistic grammar. En K. Aijmer y B. Altenberg, eds. *English corpus linguistics: Studies in honour of Jan Svartvik*. London: Longman, pp. 30–43.

ISSN: 1139-8736 https://infoling.org/elies/

Hanks, Peter; Pustejovsky, James. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée*, 10.2: 63–82.
Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45.1: 31–80.
Heylen, Kris; Tummers, José; Geeraerts, Dirk. 2008. Methodological issues in corpus-based Cognitive Linguistics. En G. Kristiansen, ed. *Cognitive Sociolinguistics: Language variation, cultural models, social systems*. Berlin: Mouton de Gruyter, pp. 91–128.
Hunston, Susan; Francis, Gill. 1998. Verbs observed: A corpus-driven pedagogic Grammar. *Applied Linguistics* 19.1: 45–72.
Hunston, Susan; Francis, Gill. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
Ide, Nancy; Pustejovsky, James, eds. 2017. *Handbook of Linguistic annotation*. Dordrecht: Springer.
Johnston, Trevor. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15.1: 106–131.
Kipp, Michael; Martin, Jean-Claude; Paggio, Patrizia; Heylen, Dirk, eds. 2009. *Multimodal corpora: From models of natural interaction to systems and applications*. Berlin: Springer.
Knight, Dawn; Adolphs, Svenja. 2020. Multimodal corpora. En M. Paquot y S. Gries, eds. *A practical handbook of Corpus Linguistics*. Springer, pp. 353–370.
Labov, William. 1972. The study of language in its social context. En J.A. Fishman, ed. *Advances in the Sociology of language*, v. 1. The Hague: Mouton, pp. 152–216.
Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar*, Volume 1: *Theoretical Prerequisites*. Stanford: Stanford University Press.
Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. En J. Svartvik, ed. *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, pp. 105–122.
Manning, Christopher. 2003. Probabilistic syntax. En R. Bod, J. Hay y S. Jannedy, eds. *Probabilistic Linguistics*. Cambridge: MIT Press, pp. 289–341.
Martí Antonín, Antonia M. 2018. Modelos de semántica distribucional. En M. Diaz-Ferro et al., eds. *Actas do XIII Congreso Internacional de Lingüistica Xeral*. Universidade de Vigo, pp. 16–22.
McEnery, Tony; Xiao, Richard; Tono, Yukio. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.
Molina Salinas, Claudio; Sierra Martínez, Gerardo. 2015. Hacia una normalización de la frecuencia de los corpus CREA y CORDE. *Revista signos* 48.89: 307–331.
Mukherjee, Joybrato. 2004. Corpus data in a usage-based cognitive grammar. En K. Aijmer y B. Altenberg, eds. *Advances in Corpus Linguistics*. Amsterdam: Brill Rodopi, pp. 83–100.
Pérez, Ania; García-Miguel, José M.; Cabeza, Carmen. 2019. Anotación de corpus para o estudo da expresión gramatical de eventos: notas sobre o deseño do proxecto RADIS. *Sensos-e* 6.1: 40–61.
Rojo, Guillermo. 2016. Los corpus textuales del español. En J. Gutiérrez-Rexach, ed. *Enciclopedia de Lingüística Hispánica*. Oxford: Routledge, pp. 285–296.
Rojo, Guillermo. 2021. *Introducción a la lingüística de corpus en español*. London: Routledge.

ISSN: 1139-8736 https://infoling.org/elies/

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, John. 2004. *Trust the text: Language, corpus and discourse*. Londres: Routledge.
Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology*. Berlin: Language Science Press.
Stefanowitsch, Anatol; Gries, Stefan. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8.2: 209–243.
Stubbs, Michael. 1996. *Text and corpus analysis*. Oxford: Blackwell.
Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.
Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins.
Torruella, Joan. 2017. *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística.* Frankfurt am Main: Peter Lang.
Torruella, Joan; Llisterri, Joaquim. 1999. Diseño de corpus textuales y orales. En J.M. Blecua, G. Clavería, C. Sánchez y J. Torruella, eds. *Filología e informática: Nuevas tecnologías en los estudios filológicos*. Barcelona: UAB / Ed. Milenio, pp. 45–81.
Tummers, Jose; Heylen, Kris; Geeraerts, Dirk. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1.2: 225–261.
Valenzuela, Javier. 2022. El big data en los estudios del lenguaje. *Estudios de Lingüística del Español* 45: 241–260.

ISSN: 1139-8736 https://infoling.org/elies/