Parte I

Tipología de corpus para el español

1

Constitución de corpus crecientes del español

(Creating a monitor corpus of Spanish)

Mark Davies y Giovanni Parodi

1. Introducción

En los últimos cinco o diez años disponemos de un número importante de grandes corpus del español (para ser considerados así deben tener un tamaño de, al menos, 100 millones de palabras): Corpus Sketch Engine, Corpus de la Web, Corpus del Español y Corpus de la Real Academia Española. Tal como señala Davies refiriéndose a los corpus en inglés, existe una estrecha relación entre el tamaño del corpus y el rango de fenómenos lingüísticos que se pueden estudiar. Un corpus con uno o dos millones de palabras es lo más aconsejable para estudiar fenómenos de alta frecuencia, tales como marcadores del discurso, preposiciones o construcciones sintácticas frecuentes (pasivas, perfectivas o progresivas). Sin embargo, los corpus pequeños a menudo no son los más idóneos para investigaciones detalladas de léxico, morfología, preferencias colocacionales o para construcciones sintácticas de mediana o baja frecuencia (tales como los complementos verbales). Además, los grandes corpus -si poseen el tipo correcto de arquitectura e interfaz- también pueden proveer información valiosa acerca de la variación entre géneros del discurso, dialectos y períodos temporales. En este capítulo, primeramente, presentaremos (Sección 1) los corpus que constituirán el foco de nuestra discusión, refiriéndonos, entre otros, a su tamaño, composición, período de tiempo y dialectos. En la Sección 2 abordaremos algunos aspectos básicos de la anotación de estos grandes corpus —principalmente referidos al etiquetaje de categorías gramaticales y lematización— y también de la precisión de las anotaciones en los diferentes corpus. La Sección 3 mostrará en mayor detalle diversos aspectos de la funcionalidad del corpus y discutiremos cómo el tamaño y composición del corpus impacta en cada uno de estos. Ello incluye: (i) listas de frecuencia, que incluye cadenas de palabras y listas completas de palabras; (ii) construcciones gramaticales (con categorías gramaticales y lemas); y (iii) colocaciones (examinando el significado y uso de las palabras).

La Sección 4 abordará cuán grande ha de ser el corpus para el estudio de los géneros, los dialectos y el desarrollo histórico. La Sección 5 incluirá muy brevemente los corpus paralelos de gran tamaño que se pueden emplear para comparar el uso y encontrar traducciones en dos lenguas y por último se ofrecen algunos comentarios de cierre y proyecciones.

DOI: 10.4324/9780429329296-3

Palabras clave: grandes corpus; Sketch Engine; Corpus de la Web; Corpus del Español; CREA, CORPES

The last five or ten years have seen the release of several very large corpora of Spanish (defined here as corpora with at least 100 million words in size): corpora from Sketch Engine, corpora from the web, the Corpus del Español, and the Real Academia Española. As Davies discusses (regarding English corpora), there is a fairly close relation between corpus size and the range of linguistic phenomena that can be explored. A one- or two-million-word corpus is best for very high frequency phenomena, such as discourse markers, prepositions, or frequent syntactic constructions (passive, perfect, or progressive). But small corpora such as these are often quite inadequate for detailed investigation of lexis, morphology, collocational preferences or medium and low-frequency syntactic constructions (such as verbal complementation). In addition, very large corpora—if they have the right type of architecture and interface—can also provide valuable insight into variation between genres, dialects, and time periods. In this chapter, we will first outline (in Section 1) the corpora that are the focus of our discussion – by discussing their size, composition, time periods, dialects, and so on. Section 2 will discuss some basic issues of annotation for these large corpora—mainly part-of-speech tagging and lemmatization—as well as the accuracy of annotation in the different corpora. Section 3 will provide a more detailed examination of several different aspects of corpus functionality and will discuss how the size and composition of the corpora impact each of these. These will include: (i) frequency lists, including matching strings and full wordlists; (ii) grammatical constructions (with part-of-speech and lemma); and (iii) collocates (to examine word meaning and usage).

Section 4 will consider how large corpora can be used to examine genre-based, dialect and historical. Section 5 considers very briefly the topic of large, parallel corpora that can be used to compare usage and find translations in two languages, and then offers some concluding remarks and projections.

Keywords: large corpus; Sketch Engine; Web Corpus; Corpus del Español; CREA; CORPES

2. Estado de la cuestión: colecciones de corpus

Sketch Engine (www.sketchengine.eu; Kilgarriff *et al*. 2014) contiene la mejor colección de los corpus crecientes para una amplia variedad de lenguas. Sketch Engine posee corpus de, al menos, mil millones de palabras para más de veinte lenguas diferentes. El corpus principal del español se denomina [esTenTen18] y una descripción de ese corpus se puede encontrar en www.sketchengine.eu/estenten-spanish-corpus/.

EsTenTen18 actualmente contiene información de cerca de 18 mil millones de palabras, lo que lo hace el corpus estructurado más grande del español. 79% de su base de datos proviene de Latinoamérica mientras que el 21% viene de España. Los cinco países que proveen más corpus son Argentina, España, México, Chile y Colombia. Además del corpus principal esTenTen18, también está disponible un reciente corpus con marcas temporales (*Timestamped*) (www.sketchengine.eu/timestamped-spanish-corpus/), el cual es continuamente actualizado. Desde finales de 2019, este corpus tiene alrededor de 8 mil millones de palabras y está creciendo a un ritmo de 150 millones de palabras cada mes.

Así como veremos, además de ser muy grande, los corpus de Sketch Engine también tienen una avanzada y robusta arquitectura e interfaz. Los corpus disponibles poseen un amplio rango de búsquedas y son particularmente conocidas por la facilidad que permiten a los usuarios examinar las colocaciones de una palabra o hacer comparaciones entre palabras. Estos corpus

también tienen una gran capacidad para crear "corpus virtuales" (subcorpus) y luego investigar dentro de estos. Los corpus son etiquetados con FreeLing, el que será abordado en la Sección 2.

Muy relacionado con el Corpus Sketch Engine, tenemos el Corpus de la Web (https://corporafromtheweb.org) que incluye corpus del español, inglés, francés, alemán, danés y sueco. El Corpus de la Web está relacionado con el Corpus Sketch Engine en el sentido que la arquitectura subyacente para el Corpus de la Web es "No Sketch Engine", lo cual de alguna manera es una versión limitada de la arquitectura "Sketch Engine". Sin embargo, el Corpus de la Web tiene muy bien integradas otras herramientas como el R Studio. El Corpus del español (http://corporafromtheweb.org/escow14/#more-76) tiene información de cerca de siete mil millones de palabras las que fueron recolectadas entre los años 2011 y 2014 (tal como para el Sketch Engine) y fue etiquetado y lematizado con el etiquetador FreeLing.

Existe un número considerable de corpus disponibles en el sitio web Corpus del Español (www.corpusdelespanol.org). El más ampliamente usado de ellos es el Corpus Web-Dialects, el cual está disponible desde el 2017. Tiene información acerca de dos mil millones de palabras, recolectadas de 21 países de habla hispana en 2014 (78% de Latinoamérica y 22% de España). Aproximadamente, 50% de la información proviene de *blogs*, lo que significa que los investigadores pueden estudiar rasgos altamente informales de cada lengua. La estructura y arquitectura de este corpus del español es muy similar a la del corpus del inglés GlobWbE (Global Web-based English) que se presenta en Davies y Fuchs (2015).

Un segundo gran corpus es el Corpus NOW (News on the Web). Tiene sobre siete mil millones de palabras y es tres o cuatro veces tan grande como el Corpus Web/Dialects. A finales del 2019, estaba creciendo cada mes con cerca de 100 millones de palabras y tiene la misma arquitectura e interfaz del corpus English NOW. Debido a los nuevos textos que se han agregado al corpus cada mes desde 2012 a 2019, permite a los investigadores estudiar cuidadosamente los cambios más recientes de la lengua.

El corpus más pequeño es el Corpus Historical/Genres, que contiene 100 millones de palabras entre los años 1200 al 1900. Para el Corpus Español Moderno (1800–1900), está dividido en partes similares en géneros orales, de ficción, periodísticos y académicos, lo que permite a los investigadores comparar todos estos géneros. Debido a que este corpus ha sido previamente discutido en otra parte (ver Davies 2008, 2010 y 2018b) y también debido a que la parte "moderna" del corpus (1800–1900) solo tiene 40 millones de palabras, no abordaremos este corpus en este capítulo.

Por último, existen otros recursos que están basados en una gran cantidad de información del Corpus Web/Dialects. Ello incluye información de texto completo que puede ser descargable y empleada en los computadores de los propios investigadores. Además, en el WordAndPhrase.info/Span, profesores y estudiantes pueden examinar 40 000 lemas en el Corpus Web/Dialects y observar abundante información de cada palabra —frecuencia, distribución de género, sinónimos, colocaciones, tópicos relacionados y líneas de concordancia. También se puede ingresar textos completos y analizarlos generando listas de palabras desde el mismo texto.

En las secciones 3 y 4, discutiremos el modo en que la arquitectura e interfaz del corpus del español permite investigar por palabras, frases y patrones; complejas cadenas que involucran categorías gramaticales, lemas, sinónimos y lista de palabras personalizadas; colocaciones; y comparaciones entre géneros, dialectos y periodos de tiempo. La completa arquitectura e interfaz están orientadas a examinar la variación —variación entre dialectos en el caso del Corpus Web/Dialects y a través tiempo en el caso del Corpus NOW. Para una discusión más profunda de la arquitectura e interfaz de este corpus se recomienda consultar Davies (2017, 2018a).

Los corpus más pequeños de estos grandes corpus son los corpus de la Real Academia Española (RAE). El más reciente de estos Corpus del Español del Siglo XXI o CORPES (www.rae.es/recursos/banco-de-datos/corpes-xxi). El CORPES contiene 175 millones de palabras recogidas entre el 2001 y el 2015, 65% provienen de textos de América Latina y 35% de España. Así como será discutido en más detalle en la Sección 3, tenemos allí una buena cobertura de textos de diferentes dialectos. Desafortunadamente, no es posible comparar entre los diferentes dialectos, excepto al nivel individual de palabra. Además, como veremos en la Sección 3, el corpus es muy limitado en términos de colocaciones, concordancias, listas de frecuencia y otros tipos de búsquedas más allá de palabras simples o frases exactas.

Así como para el Corpus del Español, también existen corpus de la RAE que fueron creados 20 o 25 años atrás. Los dos corpus principales son el Corpus de Referencia del Español Actual (CREA) (www.rae.es/recursos/banco-de-datos/crea), el cual contiene 160 millones de palabras recogidas desde finales de 1900 hasta inicios del 2000, y el Corpus Diacrónico del Español (CORDE) el cual contiene 125 millones de palabras correspondientes a periodos anteriores del español. Tal como discutiremos en más detalle en la Sección 3, ambos corpus permiten observar la frecuencia de una palabra o frase exacta por género del discurso o periodo de tiempo. Pero ambos corpus son extremadamente limitados en términos de búsquedas de colocaciones, concordancias, listas de frecuencias o cualquier otra búsqueda que no sea de solo palabras o frases exactas.

3. Anotaciones de corpus

Tal como ya lo hemos discutido, existen varios corpus del español en Sketch Engine, Corpus de la Web y Corpus del Español, los cuales contienen más de mil millones de palabras. El Corpus esTenTen18 de Sketch Engine tiene un tamaño cercano a 18 mil millones de palabras. Sin embargo, el tamaño no lo es todo. Una vez que el corpus se ha creado, se anotan las categorías gramaticales (e.g., un *trabajo = noun, yo trabajo = verb)* y los lemas e.g., *dice, dijo y diremos* son todas formas del lema *decir*). Mientras hoy en día es muy fácil crear un gran corpus de la web para cualquier lengua, es mucho más difícil anotar estos corpus correcta y adecuadamente. Y sin buenas anotaciones, el corpus es prácticamente inútil, al menos para ciertos propósitos.

Por ejemplo, generamos desde el Sketch Engine una lista de todos los lemas que ocurren 20 veces o más en un corpus. Para el propósito de este capítulo, nos enfocaremos en aquellos verbos que comienzan con [s-] (ser, salir, etc.; la lista completa se encuentra en www.corpusdelespanol.org/files/se_span.xlsx). Es posible darse cuenta de que la lista de los lemas se inició correctamente. Por ejemplo, los diez primeros verbos son: *ser, saber, seguir, salir, señalar, sentir, servir, solicitar, suponer, sacar*, los cuales son todos de alta frecuencia en español. Así que, para varias palabras frecuentes, parece que el etiquetador FreeLing lo hizo muy bien. Bajo la palabra #1.000, encontramos los siguientes lemas uno después del otro: *satisfacer, siempre, saco, simone, somos, siguió, sperar, sustituído, supply, safó, sardinada, subiamos, subway, sobrescribe, soñabamos, sección, subredondear, santalucía, scripta, scuba, seleccionada, sosteían, surfea, sarpado, satisfacción, sorprendido, sugiere, semibatir, september, seva*. Virtualmente ninguno de estos verbos son auténticos lemas. Son formas (o formas cercanas) de lemas, pero no son lemas verdaderos (*somos, soñábamos, sugiere, substituído, subíamos, sostenían*) o ellos son de otro idioma (*supply, subway, scuba, September*) o ellos son extraños (*simone, santalucia, seva*).

Y esto es cerca del inicio de la lista, donde presumiblemente se podría haber corregido los primeros 1 000 verbos, si se hubiera sabido español o si se hubieran tomado el tiempo de

corregir los resultados del etiquetador morfosintáctico y el lematizador. Las cosas se vuelven mucho más extrañas cuando se avanza en el listado. Por ejemplo, cerca del verbo #3 200 encontramos los siguientes lemas: *salienron, salomé, sangree, scarce, scrooge, sdfr, sebita, seeeeeeeeeee, separació, serásn, sexan, shay, shúper, silicone, simos, siome, ske, sommer, sorcerer, spaña, swear, self-care*. Ninguno de estos son lemas verbales y ninguno ha sido corregido de ninguna manera.

Dado que los resultados del etiquetador morfosintáctico y el lematizador resulta tan problemático, probablemente existirá un número de casos en los cuales los resultados de nuestra búsqueda estarían seriamente comprometidos por la calidad de estos datos. Las listas de frecuencias de palabras serían sospechosas; las listas de colocaciones serían confusas y las búsquedas sintácticas que involucran verbos incluirían del mismo modo información distorsionada. Ciertamente no existe un etiquetador morfosintáctico o lematizador que sea completamente certero y no sería justo focalizarse solo en aquellas entradas problemáticas. Pero cuando existen tantos errores no corregidos y altamente frecuentes se tienen razones para una justa preocupación.

Al menos hasta donde sabemos, el único corpus grande del español con más de un billón de palabras es el Corpus del Español. El marcaje y lematización original de los dos mil millones de palabras del Corpus Web/Dialects fue realizado con el etiquetador morfosintáctico Palabras de Eckhard Bick, el cual está basado en el altamente preciso etiquetador Palavras (ver Bick 2000) para el portugués. Incluso con este etiquetador se producen muchos errores, algunos de ellos con alta frecuencia. Por lo tanto, se invirtieron tres a cuatro meses con cientos de horas corrigiendo el resultado del etiquetador *Palabras*. Debido a que esos dos mil millones de palabras estaban almacenadas en una base de datos altamente estructurada, se pudo corregir esos errores eficientemente.

Por ejemplo, se pudo rápidamente buscar y corregir todos los verbos que tenían una o dos formas (e.g., el erróneo verbo aguar para la frecuente palabra agua). Otro ejemplo es que se calculó la frecuencia de [artículo + x + preposición/conjunción] para todos los supuestos sustantivos, tales como: el hombre que, unos casos de, una casa en. Cualquier palabra x que no ocurría muy frecuentemente en este marco sintáctico era sospechosa y fue examinada individualmente. En un tercer ejemplo, se comparó la lista de los primeros 40 000 supuestos lemas con cinco diccionarios en línea del español (los cuales se habían trasladado y estaban ahora en una base de datos relacional). Cualquier lema que no estuviera en estos diccionarios fue catalogado de sospechoso y fue examinado manualmente (esto habría eliminado errores tales como: *scrooge, sdfr, sebita, seeeeeeeeeee*, listados más arriba para el Corpus Sketch Engine*)*. Estos son solo tres de cientos de diferentes maneras en las cuales se revisó la precisión de las etiquetas y los lemas en el Corpus Web/Dialect y por esto es que sostenemos que es el único grande y preciso corpus del español.

4. Consideraciones metodológicas: funcionalidad de los grandes corpus

Los corpus permiten generar listas de frecuencias de palabras, encontrar la frecuencia de palabras frases y n-gramas (cadenas con N número de palabras), investigar construcciones sintácticas, realizar otras búsquedas que aborden información semántica y extraer las colocaciones de palabras y frases. Todo esto puede ofrecer una mirada muy útil respecto del uso y el significado de las palabras. En las secciones siguientes mostraremos como todos estos tipos de búsquedas operan mejor con grandes y bien diseñados corpus.

4.1. Listas de palabras

El Corpus Sketch Engine, el Corpus de la Web y el Corpus del Español todos permiten a los usuarios crear listas de palabras desde la información del corpus. En el caso del Corpus del Español se ha diseñado un completo sitio web con una lista de aproximadamente 40.000 palabras. Por su parte, el corpus WordAndPhrase-Spanish (www.wordandphrase.info/span/) permite a los usuarios explorar las 40.000 primeras palabras del español y observar un amplio rango de información de ellas, incluyendo definiciones, frecuencias, dispersiones, sinónimos, colocaciones y tópicos relacionados (palabras que coocurren en cualquier parte de la página web), distribuciones dialectales y líneas de concordancia de KWIC (del inglés *Key Word in Context*). Los usuarios pueden también moverse continuamente de una palabra a otra; casi todo en cada una de las 40.000 páginas de inicio para cada palabra es accesible. Además de este sitio web, también existe un diccionario impreso de frecuencias (Davies y Davies 2017) que está basado en la información del corpus y está orientado a aprendientes y profesores de español.

Mientras más grande el corpus, existirán mayores ocurrencias de palabras de baja frecuencia. Por ejemplo, en los 20 millones de palabras del Corpus del Español (Historical/Genres) existen solo 6.500 palabras con una frecuencia de 150 ocurrencias o más. Ejemplos de lemas de palabras con una frecuencia de 150 ocurrencias son: (adjetivo) *acreditado, desordenado, destructivo, digestivo, director, irreversible, próspero, quemado*; (sustantivo) *compasión, crepúsculo, creyeras, nte, descentralización, emigración, haitiano, ovario, prudencia, tejado, tobillo, yegua*; and (verbo) a*cechar, atropellar, azotar, consignar, empeorar, superponer*. Debido al pequeño tamaño del corpus, si quisiéramos una lista de 40.000 palabras, las palabras menos frecuentes ocurrirían entre cuatro y cinco veces cada una. En ese punto, eso sería probablemente solo "ruido" -si uno o dos de los textos en el corpus fuera diferente, entonces aquellas palabras no estarían en nuestra lista.

Con un corpus de dos mil millones de palabras como el Web/Dialects (100 veces más grande que Historical/Genres), la lista es mucho más confiable para palabras de más baja frecuencia. Por ejemplo, las siguientes son palabras del final de la lista de 40.000 palabras, e incluso todas ellas ocurren cerca de 145 veces o más que es mucho más que meramente "ruido"): (adjetivo) *aislacionista, corvino, nugatorio, mesquino, pérsico, micrométrico, cesionario*; (sustantivo) *rumbero, pelafustán, gocho, acristalamiento, acromegalia, comercialismo, borreguismo*; (adverbio) *comprensivamente*; (verbo) *intersecar, puncionar*. Pero una vez más, solo tener muchas ocurrencias no es suficiente. La lista requiere ser cuidadosamente revisada como se dijo anteriormente.

Con un corpus más grande, se puede llevar a cabo mejores análisis de la morfología del español. Por ejemplo, en los 20 millones de palabras del Corpus Historical/Genres, existen sólo alrededor de 45 palabras terminadas en **azo* (sufijo que alude a un movimiento o golpe con un objeto) que ocurren 20 veces o más; y las más bajas frecuencias son *arañazo, terrazo, espaldarazo, golazo, pinchazo, pelotazo, escopetazo, campanillazo, gustazo, trancazo*. Tenemos una información mucho más rica con los dos mil millones de palabras del Corpus Web/Dialects donde hay más de 460 palabras que ocurren al menos 20 veces, incluyendo las de baja frecuencia: *morongazo, postazo, teclazo, pollazo, autogolazo, libretazo, arcabuzazo, platanazo, trolazo, inventazo, peloterazo, guadañazo, potazo, fregadazo, cuponazo*. Obviamente con más de 20 veces en muchos tipos (formas distintas) se puede investigar el fenómeno de la creatividad léxica de mucho mejor manera que con un corpus pequeño.

4.2. Secuencias de palabras y n-gramas

Los grandes corpus también pueden ser útiles para generar n-gramas, o palabras que contengan N número de palabras en una cadena. Estos n-gramas se emplean en Procesamiento del Lenguaje Natural para ayudar a los computadores a reconocer y procesar cadenas de palabras. Por ejemplo, un programa computacional para un teléfono móvil puede pensar que existe una cadena de tres palabras cuya palabra media es *cuenta*. Pero para procesar adecuadamente esta cadena, se necesita saber la frecuencia de tres cadenas de palabras de la forma [X cuenta Y]. Los corpus pueden proveer esta información.

Por ejemplo, existen cerca de 2.000 trigramas en el Corpus Web/Dialects que arrojan una frecuencia de al menos 30 ocurrencias incluyendo 30 casos de baja frecuencia: *darán cuenta por, dé cuenta a, de cuenta ya, daras cuenta, darás cuenta, capital cuenta con, apenas cuenta, app cuenta con, en cuenta tan, editorial cuenta con, doy cuenta ya*. En el Corpus Historical/Genres, existen solo 64 trigramas con [X cuenta Y] que ocurren al menos 30 veces. En otras palabras, con el corpus de dos mil millones de palabras existen cerca de 60 veces más trigramas que ocurren por lo menos 30 veces, lo que significa que existe una probabilidad mucho más grande de que la información del corpus sea útil en ayudar al programa computacional a analizar el español.

Los n-gramas pueden ser fácilmente generados con el Corpus del Español, y un tanto más difícilmente con el Sketch Engine y el Corpus de la Web. Con el Corpus del Español, los usuarios simplemente ingresan la cadena de palabras [* cuenta] para encontrar trigramas con *cuenta* en posición intermedia y el corpus generaría resultados desde los dos mil millones de palabras en menos de un segundo. La velocidad y facilidad en la extracción de los n-gramas en el Corpus del Español se debe a la arquitectura subyacente al corpus, la cual almacena el corpus total (en el caso del Corpus Web/Dialects) como dos mil millones de filas consecutivas con 21 palabras de contexto (la palabra núcleo con diez palabras de contexto a la izquierda y diez palabras de contexto a la derecha).

4.3. Construcciones gramaticales

Otra ventaja de los grandes corpus es la posibilidad de observar un amplio rango de construcciones sintácticas. Como Davies (2015) explica, con corpus pequeños de 10 a 20 millones de palabras estamos limitados principalmente a mirar las construcciones de alta frecuencia, tales como las progresivas (*estaba pensando*), las pasivas (*fue pintado ayer*) y las perfectivas (*ya han salido*). Con los grandes corpus, sin embargo, también podemos observar las construcciones de mediana y baja frecuencia tales como los complementos verbales.

Como un ejemplo concreto de información sintáctica mucho más rica que está disponible en los corpus más grandes, se puede considerar, por ejemplo, la frecuencia total de ciertas construcciones causales en dos corpus a partir del Corpus del Español. Si se busca le|le HACER _vr se (ya sea *le* o *les* + cualquier forma de *hacer* + infinitivo + *se*, e.g., *les hace sentirse, le haga sentirse, le hicieron ganarse, les hizo darse, le hagan decidirse*). En los 40 millones de información léxica disponible del Corpus Historical/Genres, existen solo 134 ocurrencias de esa construcción. Por otro lado, en el Corpus Web/Dialects de dos mil millones de palabras existen 3272 ocurrencias – casi 25 veces más información. Como se aprecia, para construcciones sintácticas de baja frecuencia los grandes corpus proveen una información mucho más rica para el análisis.

Probablemente, los más limitados de los cuatro grandes corpus en términos de búsquedas sintácticas son los corpus de la RAE. Si quisiéramos investigar la variación en el uso de una

partícula reflexiva como *se* en una construcción causal, por ejemplo, *le hace sentirse mal en su pensamiento*. Para el propósito de este análisis esto se compondría de algo así como: [pronombre + una forma de hacer + un infinitivo + *se*]. En el CORPES, los usuarios deben ingresar un término por ejemplo (formas de *hacer*) luego "Proximidad" y luego otro término, luego "Proximidad" y así sucesivamente tal como se aprecia en la Figura 1.1 y en los resultados desplegados en la Figura 1.2.

Las búsquedas sintácticas son mucho más fáciles de realizar con el Corpus Sketch Engine y con el Corpus de la Web, ellos emplean el poderoso *Corpus Query Language (CQL)*. Existen pequeñas diferencias en las búsquedas sintácticas, de modo que en el Corpus Sketch Engine sería algo como: [palabra = "le|les"] [lema = "hacer"] [palabra = ".*rse"] que brindaría 32 458 ocurrencias), mientras que en el Corpus de la Web sería: [palabra = "le|les"] [lema = "hacer"] [tag = "V.*"] [palabra = "se"] lo que obtendría 6 197 ocurrencias). En ambos casos, el corpus primero entrega todas las líneas de concordancia coincidentes y luego el usuario puede ver un resumen de la frecuencia de las cadenas obtenidas (Figura 1.3).

El Corpus del Español emplea probablemente la búsqueda sintáctica más poderosa; las búsquedas pueden incluir cualquier combinación de palabras, subcadenas, categorías gramaticales, lemas, sinónimos, o listas de palabras personalizadas. Por ejemplo, una búsqueda de verbo causal sería: [le|les HACER VERB se]. El corpus primero muestra la frecuencia de las cadenas y luego las líneas de concordancia para las cadenas seleccionadas tal como se muestra en las siguientes Figura 1.4 y Figura 1.5.

Además de palabras, categorías gramaticales y lemas (con CQL), el Corpus del Español también permite búsquedas "de base semántica" que incluyen sinónimos y listas de palabras personalizadas como en: =ARGUMENTO =LÓGICO. Esto buscaría cualquier forma de un sinónimo de *argumento* seguido de cualquier forma de un sinónimo de *lógico*. Ello entregaría cadenas del tipo*: juicio justo, explicación lógica, razonamiento lógico, explicación racional, conclusión lógica, razón natural, razón lógica, argumentos racionales*.

Otro ejemplo podría ser [PRON PONER * @ROPA @COLORES]. En este caso @ indica unas listas de palabras personalizadas que el usuario ha creado y de este modo esto sería interpretado como "pronombre + cualquier forma de *poner* + cualquier palabra +una palabra en la lista de ropa + una palabra en la lista de colores". La consulta toma menos de un segundo

para buscar la información en dos mil millones de palabras y el resultado sería el que se presenta en la Figura 1.6.

4.4. Colocaciones

Las colocaciones son un listado de palabras adyacentes que coocurren con una palabra determinada y ellas se constituyen en herramientas poderosas para investigar el significado y el uso de una palabra. Tal como señaló Firth (1957, 11), "you shall know a word by the company it keeps".

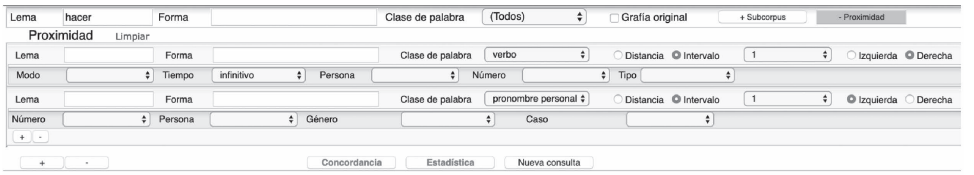


Figura 1.1 Búsquedas sintácticas en el CORPES (RAE): búsqueda de forma.

41.188 casos en 15.716 documentos.

REF.	(Clasificación, país)	CONCORDANCIA		Ordenar por: Año ascendente / sin criterio
1	2001 Esp.	beben coca-cola y comen hot-dogs son simples invitados. Sus "dueños" -¡y bien que lo	hacen	ver!- son aristócratas de Mayfair, brigadiers de Kent y caporales de Yorkshire,
2	2001 Esp.	una mínima referencia en el "Infierno" de Dante a un tramposo florentino que se	hizo	pasar por un vecino para heredar su fortuna.
3	2001 Esp.	segunda lectura aportaría, podría pensarse, poco más. Un error, claro, porque No me	hagas	pensar está escrito con tanta inteligencia y humor, destilando con sencillez la
4	2001 Esp.	Para explicar de qué va No me	hagas	pensar voy a tomar del texto un ejemplo evidente para cualquiera que haya intentado
5	2001 Esp.	precisamente es la máxima de Steve Krug, su gran ley de la usabilidad en la web: ¡No me	hagas	pensar! Su postura es muy simple: si un usuario llega a una página web, lo que
6	2001 Esp.	No me	hagas	pensar es un delgado volumen, poco más de doscientas páginas, que sin retórica
7	2001 Bol.	plato tradicional y popular de Combaya, se lo prepara del conejo nativo y se lo	hace	cocer en ají amarillo molido, con harto ajo y aderezos. Es vianda que caracteriza
8	2001 Méx.	hombre las estaba manteniendo con sangre humana, y no sangre de venado, como les	hacía	creer, las Pléyades se enojaron mucho y se marcharon al cielo.
9	2001 Méx.	se acerque a la hembra, pero el perro amarillo lo mira ferozmente a los ojos y lo	hace	retirarse.
10	2001 Méx.	atravesaron el desierto de Chihuahua y cruzaron la frontera norteamericana. En Chicago se	hicieron	llamar Twin y Twiligth. No formaron parte de ninguna pandilla, ellos mismos fueron

Figura 1.2 Búsqueda sintáctica en el CORPES (RAE): resultados.

Query le|les, hacer, V.*, se **6,197** (0.90 per million)

Page 1 of 310 Go Next | Last

6e8b07cf1c08f00e244b0c173fa634a3055d	empeorado . Espero que pronto la rentabilidad	le haga unir se	a nuestros clientes . De cualquier forma a
ce786362982ef862908c5ca52fe81f17251a	fue muy enfermizo , y cuando no es una alergia que	le hace rascar se	y sufrir los picores , es una indigestión o un
43d8a040d4730216a13db0ea7c1305217524	interesantes y hasta humorísticas , lo que	le hizo interesar se	cada vez más en las materias . Como resultado ,
bbbfb3c8e9e5372e51fdfed4feb5d9891e87	con todos los demás subsistemas . Esta doctrina	le hace sentir se	importante a todo el mundo . El legislador
0755d63018c7717d0a29a2f34bffc026c1f5	, en inferioridad , recibe varios impactos que	le hacen estrellar se	en la misma sierra que nos ha acompañado a lo
6f8ffcca98ee430f1132719c544e39470056	personas que se lo merecen " , un pensamiento que	le hace sentir se	" un poco incómoda " frente a tantos que trabajan
0493d59c9b61edcf9012784588275380012d	- natural en las personas normales -	le haría dar se	cuenta a Pedro Almodóvar de la extraordinaria
32304ca5261f07e1305a4d80881e1159e1c3	. Todo ello proporciona seguridad , confianza ,	les hace sentir se	importantes y promueve valores de cooperación
0ada508d7863d951f4f377ec39475b31937c	unos días sufre terribles alucinaciones que	le hacen poner se	en la piel de Premutos , un anti-dios , un ángel
d87aff7c7dc8a4aa9dfbf96fd5bf9b456d94	el semental , que no era otro que Criss Strokes ,	le hicieron abalanzar se	sobre esa verga descomunal con esa pasión que la
2be572461cec242299fc111c4922ed695675	por el publico , su toque marginal y callejero	le hizo mantener se	siempre entre los intérpretes preferidos por

Figura 1.3 Búsqueda sintáctica en el Corpus de la Web.

1	LE HACE SENTIR SE	201	
2	LES HACE SENTIR SE	194	
3	LES HARÁ SENTIR SE	157	
4	LE HACÍA SENTIR SE	70	
5	LE HARÁ SENTIR SE	66	
6	LE HACEN SENTIR SE	65	
7	LE HAGA SENTIR SE	61	
8	LE HIZO SENTIR SE	48	
9	LE HIZO GANAR SE	44	
10	LES HACEN SENTIR SE	37	
11	LES HAGA SENTIR SE	36	
12	LE HIZO DAR SE	33	

Figura 1.4 Búsqueda sintáctica en el Corpus del Español (lista de frecuencia).

1	B PE	acusany.com	A	B	C	el tomar una píldora tres veces al día, ella encontró que el Kudzu le hizo olvidar se de fumar. Aunque no se han realizado estudios específicos sobre la
2	G ES	alcalordeloslibros.blogspot.com	A	B	C	relaciones sexuales junto a una mujer veintiún años mayor, que le da seguridad y le hace sentir se bien con sí mismo, sin tener un mundo común perc
3	G ES	arquehistoria.com	A	B	C	la calle. Los romanos que iban a las l e trinas públicas con esclavos les hacían sentar se primero a ellos en la bancada para que la piedra se calentara
4	B ES	bauldelcastillo.blogspot.com	A	B	C	un infortunio, conoce a Ally (Emilie De Ravin), una chica que le hará reconciliar se con la vida y hará que vuelva a tener sentido. Ally
5	G ES	biblioweb.sindominio.net	A	B	C	un hacker? ¿ Se han parado alguna vez a pensar qué es lo que les hace comportar se así, qué les ha convertido en lo que son? Yo
6	B PE	blogs.elcomercio.pe	A	B	C	a sus partidos para que les vean jugar, ellas matan por esos detalles, les hace sentir se especiales tener alguien así a su lado, aprovecha tu talento.
7	B ES	blogs.libertaddigital.com	A	B	C	. De momento las góticas le han hecho un hueco en su habitación. Como le hagan poner se las Doc Martens y el maquillaje de Morticia Adams junto cc
8	B ES	blogs.molinodeideas.com	A	B	C	. Posiblemente tanto el nombre del lugar como el apecto desarrapado de los cínicos les hizo ganar se el sobrenombre. Para terminar, una curiosidad r
9	B CO	bottegadivina.com	A	B	C	conocimiento instintivo de las cosas naturales que sólo los animales tienen, ese sentido que les hace recoger se ante la inminencia de una tormenta, r
10	G ES	campusvirtual.unirioja.es	A	B	C	los músicos, hasta que la contralto Karoline Ungler le cogió por los hombros y le hizo volver se para que pudiese contemplar la reacción de los asistent
11	B EC	chaulafanita.blogspot.com	A	B	C	de que me había casado así que si quieren encontrar en ese post algo que les haga decidir se debo decir les que no van por el buen camino jejeje A
12	B ES	danielfuente.blogspot.com	A	B	C	padre de la derrota sufrida en Hispania por Aníbal. Esto y sus victorias, le hicieron granjear se una lealtad inquebrantable de sus soldados y oficiales, y

Figura 1.5 Búsqueda sintáctica en el Corpus del Español (líneas de concordancia).

El tamaño de los corpus es crucial en términos de información colocacional. Como se comprende un corpus de muchos miles de millones de palabras brindará colocaciones mucho más ricas que un corpus pequeño de 20 o 100 millones de palabras. Por ejemplo, el Corpus Web/Dialects con dos mil millones de palabras del Corpus del Español tiene 725 diferentes sustantivos

1	B ES	adictosagranhermano.blogspot.com	A	B	C	Con la bata puesta Cuando Nacho dejó el deporte, el chico se puso la bata blanca y estudió auxiliar de enfermería. En la foto, le vemos con sus
2	B US	anticapitalistasburgos.blogspot.com	A	B	C	pero teniendo en nuestro punto de mira a ess dirigentes que hoy se ponen la camisa roja. La gente no es 100 % estúpida y su memoria adormecida p
3	G US	archiveofourown.org	A	B	C	Te hago un café? Si, gracias le contesto Pedro mientras se ponía el saco azul oscuro, mirándo se al espejo lo único que podía pensar era que
4	B AR	axxon.com.ar	A	B	C	el baño, se duchó, peinó, perfumó, y finalmente se puso su camisa azul, la misma camisa que Ana le había elogiado tanto en otros tiempos más
5	B GT	babosadasyhierbas.blogspot.com	A	B	C	80 años: Ni se preocupa de mirar el espejo. Simplemente se pone un sombrero rojo y sale a divertir se con el mundo. Tal vez todas debamos poner
6	B VE	blog.chavez.org.ve	A	B	C	, que me gusta mucho por cierto, el verde militar, me pondré la chaqueta verde, pero al árbitro hay que hacer le caso, uno lo que
7	B PE	blogs.elcomercio.pe	A	B	C	969335 Tenía 8 años. Se puso el vestido rojo de mamá, se calzó esos tacos negros que le bailaban en los pies
8	B PA	blogs.prensa.com	A	B	C	nunca me sentí tan disfrazada como el día de mi boda. Me puse el traje blanco, el velo, los zapatos, me miré en el espejo y dije
9	B ES	blondgirl49.blogspot.com	A	B	C	. Mi madre utiliza mucho este tipo de calzado, y hoy se puso unos pantalones negros, converse blancas y camisa blanca, e iba guapísima! Yo siemrpe c
10	G NI	carlosagaton.blogspot.com	A	B	C	Vázquez no disfraza nada, no se pone la camisa sport. Se pone la camisa blanca, corbata y actúa como un representante de los grandes gerentes. Eso
11	B CO	corazonrosa.org	A	B	C	un uniforme de gala negro, mientras que la siempre elegante Kate se puso un abrigo negro con botones de oro coronado con un broche de oro trébol
12	G ES	coreasocialista.blogspot.com	A	B	C	y yo, que ya tenía previsto que sucedería algo así, me puse un traje negro, con camisa blanca, pero sin corbata. A el desayuno, parecía

Figura 1.6 Búsqueda sintáctica en el Corpus del Español.

modifiers of "avena"			verbs with "avena" as object			verbs with "avena" as subject			"avena" and/or ...		
sativo	574	9.76	**comer**	499	4.54	**contener**	306	3.9	**cebada**	1,234	10.7
Avena sativa			comer avena			La avena contiene			cebada y avena		
integral	574	3.89	**moler**	423	7.05	**aportar**	68	2.72	**trigo**	884	9.2
de avena integral			de avena molida			La avena aporta			trigo y avena		
forrajero	315	8.78	**consumir**	337	4.47	**poseer**	61	2.11	**centeno**	787	10.47
avena forrajera			consumir avena			La avena posee			centeno y avena		
crudo	296	6	**cocer**	309	6.59	**ayudar**	34	1.5	**arroz**	592	8.17
avena cruda			de avena cocida			avena ayudan a			arroz y avena		
instantáneo	279	5.99	**tomar**	272	0.55	**engordar**	32	6.9	**maíz**	361	7.43
de avena instantánea			tomar avena			la avena engorda			avena y maíz		
coloidal	196	8.69	**incluir**	237	0.57	**proporcionar**	32	2	**miel**	360	7.72
harina de avena coloidal			incluir la avena			La avena proporciona			de avena y miel		
silvestre	170	4.66	**contener**	221	1.87	**actuar**	25	1.54	**cereal**	334	7.74
la avena silvestre			contienen avena			La avena actúa como			avena y otros cereales		

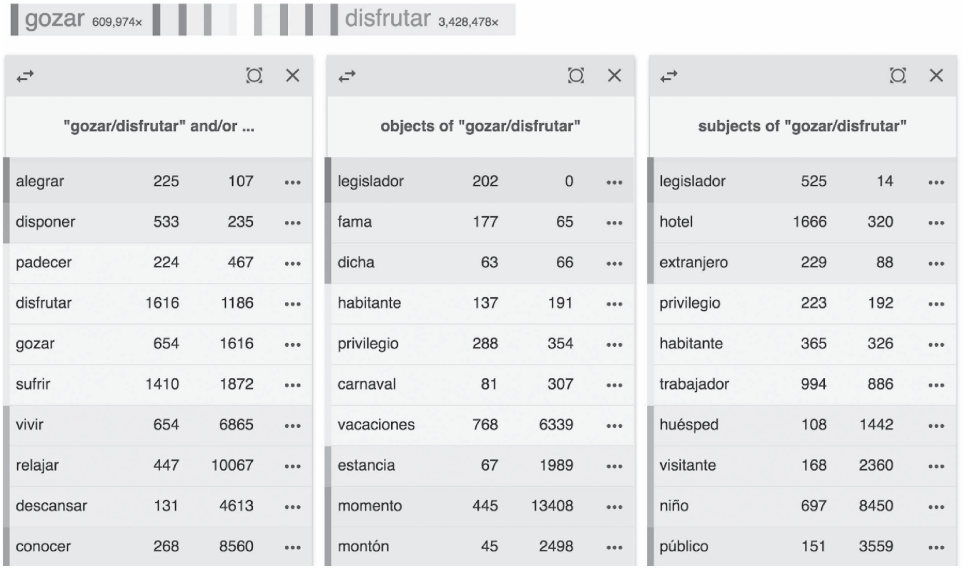
Figura 1.7 Colocaciones de *avena* en Sketch Engine (i).

que son colocaciones de *huésped*, que ocurren por lo menos cinco veces en un espacio de cuatro palabras a la izquierda y cuatro palabras a la derecha de *huésped*. Las más frecuentes serían: *casa, hotel, habitación, célula, servicio, honor, cuarto, familia, alma, viajero, trabajador, experiencia, sistema, parásito* y mucho más debajo de la lista encontramos: *transporte, ubicación, vacación, tienda, población, plato, receptor, rango, compañía, cromosoma, enfermo, entorno*. En los 40 millones de palabras del Corpus Historical/Genres, sin embargo, existen solo 21 sustantivos que ocurren al menos cinco veces como colocados de huésped.

De los cuatro grupos de los grandes corpus, el Sketch Engine entrega a los investigadores las mejores herramientas para analizar colocaciones. Como muestra la Figura 1.7, Sketch Engine provee las colocaciones en términos de función (modificador, sujeto, coordinación, etc.). Otra visualización se aprecia en la Figura 1.8, donde las colocaciones se muestran en términos de frecuencia (círculos más grandes y el puntaje de Información Mutua más cercano al centro del círculo).

En Sketch Engine, los usuarios también pueden comparar las colocaciones de dos palabras (*sketch difference*) para distinguir diferencias en su significado y en su uso. Por ejemplo, en la Figura 1.9, se muestra la comparación entre las colocaciones de *gozar* (*de*) (al inicio de cada columna) y de *disfrutar* (*de*) (al final de cada columna). Esta información podría ser útil para estudiantes, hablantes nativos de inglés, que aprenden español. En este caso, ambas palabras en inglés serían traducidas aproximadamente como "disfrutar" y los estudiantes se verían beneficiados de observar un contraste entre las dos palabras.

Figura 1.8 Colocaciones de *avena* en Sketch Engine (ii).



gozar 609,974× disfrutar 3,428,478×

"gozar/disfrutar" and/or ...			
alegrar	225	107	•••
disponer	533	235	•••
padecer	224	467	•••
disfrutar	1616	1186	•••
gozar	654	1616	•••
sufrir	1410	1872	•••
vivir	654	6865	•••
relajar	447	10067	•••
descansar	131	4613	•••
conocer	268	8560	•••

objects of "gozar/disfrutar"			
legislador	202	0	•••
fama	177	65	•••
dicha	63	66	•••
habitante	137	191	•••
privilegio	288	354	•••
carnaval	81	307	•••
vacaciones	768	6339	•••
estancia	67	1989	•••
momento	445	13408	•••
montón	45	2498	•••

subjects of "gozar/disfrutar"			
legislador	525	14	•••
hotel	1666	320	•••
extranjero	229	88	•••
privilegio	223	192	•••
habitante	365	326	•••
trabajador	994	886	•••
huésped	108	1442	•••
visitante	168	2360	•••
niño	697	8450	•••
público	151	3559	•••

Figura 1.9 Comparación de colocaciones en Sketch Engine.

Tabla 1.1 Colocaciones de *avena* en el Corpus del Español.

NOUN	*Freq.*	*MI*	*VERB*	*Freq.*	*MI*	*ADJ*	*Freq.*	*MI*
trigo	615	9.22	cebar	213	10.45	integral	380	6.71
leche	524	6.78	comer	189	3.66	rico	68	2.71
arroz	506	8.23	mezclar	126	5.29	entero	59	3.82
centeno	381	10.93	contener	121	3.47	instantáneo	52	6.28
cebada	361	11.1	salvar	106	4.17	seco	46	4.11
harina	342	8.09	preparar	86	2.86	tradicional	45	2.85
avena	292	9.95	consumir	84	3.92	cocido	43	6.97
maíz	273	7.26	recomendar	80	2.84	cereal	36	6.56
cucharada	267	8.81	agregar	60	2.58	saludable	36	3.77
copos	263	14.83	desayunar	47	6.32	sativo	35	10.38
taza	258	7.76	remojar	36	8.21	crudo	35	4.48
agua	238	3.16	adelgazar	34	5.73	quinoo	32	11.29
cereal	199	9.03	cocer	33	6.03	vegetal	32	4.11
galleta	172	8.06	arrollar	32	7.72	descremado	31	8.45

EL Corpus del Español también permite búsquedas poderosas de colocados. La Tabla 1.1 muestra para avena las colocaciones de sustantivo, verbo y adjetivo, junto con la frecuencia y el registro de Mutual Information.

El Corpus del Español también puede mostrar una comparación de dos palabras en contraste (ver Figura 1.10), para el caso de *disfrutar de* (izquierda) y *gozar de* (derecha).

En WordAndPhrase-Spanish (que se basa en el Corpus del Español), además de las colocaciones (izquierda), los usuarios también pueden observar "topics" relacionados (derecha)-palabras que coocurren en cualquier parte de la página web en los dos millones de páginas del corpus. La Figura 1.11 muestra ejemplos para *carril* (parte superior) y *conejo* (parte inferior). Los "topics" a menudo ofrecen un mejor sentido del significado de la palabra que lo que ofrecen los colocados, pero la funcionalidad "topics" solo está disponible en el Corpus del Español.

El CORPES (RAE) también puede generar una lista de colocaciones para una determinada palabra. Pero, desafortunadamente, las colocaciones se muestran ya sea por registro de asociación (Figura 1.12 para Mutual Information con *avena*) o para frecuencia bruta (Figura 1.13 para *avena*) —ninguno de los cuales por sí mismo aporta una buena imagen de la palabra.

A partir de estas informaciones, la mayoría de los usuarios que están interesados en la semántica léxica (o aquellos que enseñan a sus estudiantes el significado de palabras) probablemente se beneficiarían mucho de rasgos colocacionales en el Corpus del Español y especialmente del Corpus Sketch Engine.

5. Comparaciones entre géneros, dialectos y períodos de tiempo

Toda la discusión previa en este capítulo ha ignorado el asunto de la variación, sea esta entre géneros, entre períodos de tiempo o entre dialectos del español. Por supuesto esto no es muy realista; tal variación es la norma —ya sea léxica, morfológica, sintáctica o semántica— y se esperaría que los grandes corpus del español pudieran ayudarnos a dimensionar tal variación.

Desafortunadamente, la mayoría de los corpus de Sketch Engine y del Corpus de la Web ignoran la variación. Esto se debe en gran medida a que estos corpus son "islas" de

PALABRA 1 (P1): DISFRUTAR DE (2.99)

	PALABRA	P1	P2	FREC1/FREC2	PUNT
1	NOTICIAS	201	0	402.0	134.5
2	ACTUACIONES	125	0	250.0	83.6
3	BEBIDA	125	0	250.0	83.6
4	PASEOS	125	0	250.0	83.6
5	CARAS	106	0	212.0	70.9
6	NOTICAS	105	0	210.0	70.3
7	REVISTA	104	0	208.0	69.6
8	MOTOR	103	0	206.0	68.9
9	VÍDEOS	103	0	206.0	68.9
10	CAFÉ	205	1	205.0	68.6
11	FERIA	101	0	202.0	67.6
12	PLATILLOS	91	0	182.0	60.9
13	LETRA	80	0	160.0	53.5
14	RECETAS	77	0	154.0	51.5

PALABRA 2 (P2): GOZAR DE (0.33)

	PALABRA	P2	P1	FREC2/FREC1	SCORE
1	PRIORIDAD	63	0	126.0	376.6
2	INAMOVILIDAD	58	0	116.0	346.7
3	VALIDEZ	40	0	80.0	239.1
4	INJUSTICIA	137	2	68.5	204.7
5	PREDICAMENTO	66	1	66.0	197.3
6	INVIOLABILIDAD	27	0	54.0	161.4
7	FUERO	161	3	53.7	160.4
8	INFALIBILIDAD	26	0	52.0	155.4
9	PERSONERÍA	26	0	52.0	155.4
10	CREDIBILIDAD	243	6	40.5	121.1
11	PRIMACÍA	18	0	36.0	107.6
12	ATRIBUCIONES	17	0	34.0	101.6
13	DISCRECIONALIDAD	17	0	34.0	101.6
14	RESTABLECIMIENTO	17	0	34.0	101.6

Figura 1.10 Comparación de colocaciones en el Corpus del Español.

Figura 1.11 Colocaciones y tópicos en WordAndPhrase-Spanish.

	Clase	Freq (2)	MI	LL SIMPLE	T-SCORE
Quaker	sustantivo	12	17,51	119,75	3,46
hojuela	sustantivo	48	16,27	434,88	6,92
sativo	adjetivo	16	15,64	138,04	4,00
centeno	sustantivo	41	15,37	347,12	6,40
cebada	sustantivo	65	14,71	523,61	8,06
copo	sustantivo	32	14,27	248,46	5,65
avena	sustantivo	44	13,96	333,62	6,63
trigo	sustantivo	85	12,81	586,45	9,21
salvado	sustantivo	13	12,81	89,20	3,60
cereal	sustantivo	12	12,29	78,51	3,46
cereales	sustantivo	30	11,91	189,68	5,47

Figura 1.12 Colocaciones de *avena* en el CORPES: ranqueado por Mutual Information.

información que fueron desgajadas de la web. Es difícil comparar a través de géneros o dialectos, a menos que un investigador pueda identificar un sitio web particular (o un conjunto de sitios web) que representen adecuadamente la lengua de un género web particular o de un país determinado. Y en todos los corpus del Corpus de la Web y en la mayor parte del Corpus Sketch Engine no existe realmente una dimensión temporal, ya que estos corpus son simplemente una fotografía de la web en el momento en que esas páginas web fueron capturadas.

	Clase	Freq	MI (2)	LL SIMPLE	T-SCORE
de	preposición	668	3,7	1.241,93	23,91
,	puntuación	668	3,58	1.221,89	23,83
el	artículo	590	2,8	796,81	21,07
y	conjunción	354	3,8	611,25	17,59
.	puntuación	250	2,58	266,93	13,59
con	preposición	174	4,39	333,11	12,58
uno	cuantificador	150	3,0	177,02	10,85
en	preposición	148	2,58	143,02	10,35
??	desconocido	116	4,75	242,83	10,39
(	puntuación	104	4,95	229,56	9,90
o	conjunción	93	5,04	210,13	9,43
ser	verbo	87	2,58	79,63	7,93
trigo	sustantivo	85	12,81	586,45	9,21
)	puntuación	77	4,45	147,42	8,43

Figura 1.13 Colocaciones de *avena* en el CORPES: ranqueado por frecuencia bruta.

El Corpus del Español, por otro lado, ha sido diseñado para atender a estos tipos de variaciones. El Corpus Historical/Genres puede brindar aprendizajes valiosos respecto del cambio histórico para los últimos 800 años, del mismo modo que atiende a la variación basada en género (ver Davies 2008, 2018b). Desafortunadamente este es un corpus relativamente pequeño con solo 100 millones de palabras.

El más reciente Corpus Web/Dialects fue diseñado para ser lo suficientemente grande (con dos mil millones de palabras) con el fin de poder observar un amplio espectro de fenómenos que no podrían ser estudiados con un corpus más pequeño como el Corpus Historical/Genres (ver Sección 3). Además, sin embargo, fue diseñado para mostrar la frecuencia de cualquier palabra, frase o construcción gramatical en 20 diferentes países de habla hispana. Por ejemplo, la Figura 1.14 muestra la frecuencia de la palabra *che*! en los diferentes países (*uuuff que mala noticia che!)*, y se aprecia que es mucho más común en Uruguay y Argentina.

Por supuesto los investigadores también pueden indagar construcciones más complejas. La Figura 1.15 muestra la frecuencia para PRON-SUBJ VERB (sujeto léxico de infinitivo, e.g., *para ella entender*) en estos 20 países. El corpus muestra que la construcción es más frecuente en Puerto Rico y República Dominicana. La Figura 1.16 entrega unos pocos ejemplos de la construcción en República Dominicana con la cadena *para yo* VERB.

Más allá de buscar por palabras, frases o construcciones sintácticas específicas el Corpus del Español también puede generar una lista de aquellas palabras o frases en comparación entre dialectos. Por ejemplo, la Figura 1.17 muestra una lista de cadenas para NOUN *dulce para España (izquierda) y México (derecha):*

En términos de variación histórica el Corpus del Español-NOW (más de siete mil millones de palabras) puede mostrar la frecuencia de cualquier palabra o frase —mes a mes— desde 2012. Por ejemplo, la Figura 1.18 muestra la frecuencia de la palabra *instagramer** desde 2012:

Este corpus también puede generar una lista de todas las palabras en comparaciones entre diferentes años desde el 2012. Por ejemplo, la Figura 1.19 es una lista de las palabras terminadas en **idad* que son más comunes en 2017–2019 (izquierda) que en 2012–2015 (derecha), la Figura 1.20 compara NOUN + *de datos* en 2017–2019 vs 2012–2015 (derecha):

En términos de variación histórica, el Corpus Timestamped JSI Web del Sketch Engine permite búsquedas similares para encontrar palabras que han aumentado o decrecido desde el 2014, aunque no permite búsquedas por frase. El CREA y el CORPES son muy buenos para

SECCIÓN	TODOS	MX	GT	SV	HN	NI	CR	PA	PR	DO	CU	VE	CO	EC	BO	PE	CL	PY	UY	AR	ES
FREC	680	22	6	15	0	2	3	2	0	10	16	12	15	5	14	8	12	11	88	316	59
TAMAÑO	1950	246.0	54.3	36.5	35.1	32.4	29.6	22.3	32.2	33.7	63.2	98.2	166.5	52.4	39.4	107.3	66.2	29.8	38.7	169.4	426.6
POR MILLÓN	0.35	0.09	0.11	0.41	0.00	0.06	0.10	0.09	0.00	0.30	0.25	0.12	0.09	0.10	0.36	0.07	0.18	0.37	2.27	1.87	0.14

Figura 1.14 Frecuencia de *che!* en el Corpus del Español.

SECCIÓN	TODOS	MX	GT	SV	HN	NI	CR	PA	PR	DO	CU	VE	CO	EC	BO	PE	CL	PY	UY	AR	ES
FREC	7670	1023	279	135	112	116	126	88	254	335	219	502	726	244	77	413	289	83	108	534	1157
TAMAÑO	1950	246.0	54.3	36.5	35.1	32.4	29.6	22.3	32.2	33.7	63.2	98.2	166.5	52.4	39.4	107.3	66.2	29.8	38.7	169.4	426.6
POR MILLÓN	3.93	4.16	5.14	3.70	3.19	3.58	4.26	3.95	7.90	9.94	3.46	5.11	4.36	4.66	1.96	3.85	4.36	2.79	2.79	3.15	2.71

Figura 1.15 Frecuencia de *para* PRON VERB en el Corpus del Español.

1	B DO	laberny.net	A	B	C	la introdujo ni la sacó del programa. Mi moral es muy grande para yo caer en ese tipo de bajeza, expresó Guerrero, advirtiendo que no hablaba nunca
2	G DO	importadorandino.com	A	B	C	gabriel aquino fondeur nacido el 15 de abril del 2003. y ahora para yo saber como va el proceso no tengo el numero de caso para la residencia de
3	G DO	bienesraicesdominicana.blogspot.com	A	B	C	encuentro con que Interior y Policía me pide un acta de naturalización y que para yo sacar esa, mi mamá debe ser dominicana por nacimiento. En otras
4	B DO	artelibre.diariolibre.com	A	B	C	, ni mensaje filosófico, sino, sencillamente, que fuera una base temática para yo disfrutar la pintura. Eso me permitía entonces no aferrar me tanto a lo
5	B DO	emocionhipica.com	A	B	C	gracias a Dios. Es algo difícil de lograr para muchos jinetes profesionales y para yo tener la oportunidad de haber lo conseguido, me hace sentir muy bi
6	B DO	karmatarsis.wordpress.com	A	B	C	alcohol se combertía en un lobo feroz y mi madre no tenia el recurso para yo poder estar con ella, y yo sufri mucho, y aprendi de todo mis
7	B DO	pachaproduccion.com	A	B	C	que van hacer cuando yo me muera. Pero, que lo hagan ahora para yo ver lo. A pesar de que en su pueblo. www.ensegundos.net
8	G DO	jornadadiaria.com	A	B	C	, además que la recomendación de mi gente era que declarará de utilidad pública para yo tener acceso a la retroalimentación. agregó El ex presidente
9	B DO	sfmciudaddeljaya.com	A	B	C	la cual un poeta trata de enamorar a una campesina con métodos románticos, Para yo defender me en la zona usaba una coa, como la usaba Shaft,,

Figura 1.16 *para* PRON VERB en el Corpus del Español (concordancias).

SEC 1 (España): 426,578,900 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	PIMENTÓN DULCE	54	1	0.1	0.0	31.1
2	RECETAS DULCES	52	2	0.1	0.0	15.0
3	PUNTO DULCE	42	3	0.1	0.0	8.1
4	HISTORIA DULCE	14	1	0.0	0.0	8.1
5	MOMENTO DULCE	96	8	0.2	0.0	6.9
6	VENENO DULCE	10	1	0.0	0.0	5.8
7	JAMÓN DULCE	26	3	0.1	0.0	5.0
8	PLATOS DULCES	25	3	0.1	0.0	4.8
9	OLOR DULCE	131	18	0.3	0.1	4.2
10	RECETA DULCE	12	2	0.0	0.0	3.5
11	SANGRE DULCE	11	2	0.0	0.0	3.2

SEC 2 (México): 245,956,621 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	CAÑA DULCE	17	1	0.1	0.0	29.5
2	AMIGA DULCE	15	1	0.1	0.0	26.0
3	MUNDO DULCE	24	3	0.1	0.0	13.9
4	SORGO DULCE	19	3	0.1	0.0	11.0
5	AREPITAS DULCES	12	2	0.0	0.0	10.4
6	PAN DULCE	84	23	0.3	0.1	6.3
7	OJOS DULCES	17	5	0.1	0.0	5.9
8	VERMOUTH DULCE	13	0	0.1	0.0	5.3
9	DULCE DULCE	12	0	0.0	0.0	4.9
10	HIJA DULCE	11	0	0.0	0.0	4.5
11	SABORES DULCES	48	20	0.2	0.0	4.2

Figura 1.17 NOUN *dulce* en el Corpus del Español: España vs México.

SECCIÓN	TODOS	2012-1	2012-2	2013-1	2013-2	2014-1	2014-2	2015-1	2015-2	2016-1	2016-2	2017-1	2017-2	2018-1	2018-2	2019-1
FREC	2502	24	23	9	39	79	50	52	67	99	162	170	236	324	468	632
TAMAÑO	7200	157.0	146.8	201.0	230.9	261.9	294.3	331.1	363.6	524.2	668.4	700.5	728.9	598.1	848.6	1,102.8
POR MILLÓN	0.35	0.15	0.16	0.04	0.17	0.30	0.17	0.16	0.18	0.19	0.24	0.24	0.32	0.54	0.55	0.57
TODAS LAS SECCIONES																

Figura 1.18 Frecuencia de *instagramer** en el Corpus del Español: 2012–2019.

SEC 1 (2018-2, 2019-1, 2017-1, 201...): 3,978,921,656 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	EXHAUSTIVIDAD	58416	309	14.7	0.2	94.4
2	LIBREACTUALIDAD	378	2	0.1	0.0	94.4
3	SOBRERESPONSABILIDAD	104	1	0.0	0.0	51.9
4	EXUNIVERSIDAD	61	1	0.0	0.0	30.5
5	VELOCIDAD	209128	3740	52.6	1.9	27.9
6	ALTACALIDAD	110	2	0.0	0.0	27.5
7	BICAMERALIDAD	3703	94	0.9	0.0	19.7
8	#SOLIDARIDAD	62	2	0.0	0.0	15.5
9	ELECTIVIDAD	92	3	0.0	0.0	15.3
10	SORORIDAD	1680	57	0.4	0.0	14.7
11	AMOVILIDAD	58	2	0.0	0.0	14.5
12	PLURIANUALIDAD	250	9	0.1	0.0	13.9
13	-PUBLICIDAD	405	15	0.1	0.0	13.5
14	CUBAMOTRICIDAD	54	2	0.0	0.0	13.5
15	OPTIMALIDAD	125	5	0.0	0.0	12.5

SEC 2 (2012-1, 2012-2, 2013-1, 201...): 1,986,550,759 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	CAMUNIDAD	438	4	0.2	0.0	219.3
2	VOLATIDAD	176	14	0.1	0.0	25.2
3	IU-UNIDAD	69	6	0.0	0.0	23.0
4	LLANERIDAD	78	11	0.0	0.0	14.2
5	PUNTANIDAD	277	40	0.1	0.0	13.9
6	INDISOLUBILIDAD	246	66	0.1	0.0	7.5
7	METROSEXUALIDAD	51	15	0.0	0.0	6.8
8	RUMOROSIDAD	95	35	0.0	0.0	5.4
9	ANTI-AUSTERIDAD	53	20	0.0	0.0	5.3
10	EXPERIMENTALIDAD	97	37	0.0	0.0	5.3
11	LATOTALIDAD	236	98	0.1	0.0	4.8
12	TELE-REALIDAD	124	52	0.1	0.0	4.8
13	PROCESABILIDAD	64	28	0.0	0.0	4.6
14	CALEÑIDAD	51	25	0.0	0.0	4.1
15	MANEJABILIDAD	271	136	0.1	0.0	4.0

Figura 1.19 Palabras **idad* en el Corpus del Español: 2017–2019 vs 2012–2015.

SEC 1 (2018-2, 2019-1, 2017-1, 201...): 3,978,921,656 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	TITULAR DE DATOS	2899	11	0.7	0.0	161.8
2	ROTULADO DE DATOS	74	1	0.0	0.0	45.4
3	ENTREGA DE DATOS	5714	100	1.4	0.0	35.1
4	CATEGORÍA DE DATOS	43	1	0.0	0.0	26.4
5	MINEROS DE DATOS	60	2	0.0	0.0	18.4
6	VELOCIDADES DE DATOS	47	2	0.0	0.0	14.4
7	ENCRIPTAMIENTO DE DATOS	22	1	0.0	0.0	13.5
8	DIRECTORA DE DATOS	21	1	0.0	0.0	12.9
9	ESCÁNDALO DE DATOS	56	3	0.0	0.0	11.5
10	TRATAMIENTOS DE DATOS	262	15	0.1	0.0	10.7
11	ACTUALIZACIÓN DE DATOS	6455	497	1.6	0.2	8.0

SEC 2 (2014-1, 2014-2, 2015-1, 201...): 2,443,484,467 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SUMINISTROS DE DATOS	26	1	0.0	0.0	42.3
2	INGRESO DE DATOS	681	73	0.3	0.0	15.2
3	CANCELACIÓN DE DATOS	285	45	0.1	0.0	10.3
4	SISTEMAS DE DATOS	402	76	0.2	0.0	8.6
5	FRANQUICIA DE DATOS	30	6	0.0	0.0	8.1
6	BONO DE DATOS	116	30	0.0	0.0	6.3
7	TARIFA DE DATOS	484	126	0.2	0.0	6.3
8	BONOS DE DATOS	66	19	0.0	0.0	5.7
9	CASO DE DATOS	24	7	0.0	0.0	5.6
10	ANALISTAS DE DATOS	400	118	0.2	0.0	5.5
11	DESVANECIMIENTO DE DATOS	22	7	0.0	0.0	5.1

Figura 1.20 NOUN *de datos* en el Corpus del Español: 2017–2019 vs 2012–2015.

mostrar la frecuencia de palabras específicas a través del tiempo por género o entre dialectos. Por ejemplo, la Figura 1.21 muestra la frecuencia de la palabra lindo en diferentes años, países y géneros en el CREA.

Una desventaja del CREA, sin embargo, es que las frecuencias son frecuencias brutas y no están normalizadas considerando el tamaño de los corpus para diferentes géneros, dialectos o períodos de tiempo. Por ejemplo, si 50% del corpus es de España y 6% del corpus es de Chile, entonces presumiblemente 50% de las ocurrencias de cualquier rasgo debería provenir de España y 6% debería provenir de Chile. Si una determinada palabra ocurre 120 veces en el corpus de Chile y 500 veces en el corpus de España, esta palabra sería dos veces más frecuente (por millón de palabras en la frecuencia normalizada) en Chile que en España. No obstante, el corpus solo provee frecuencias brutas y no las frecuencias normalizadas), lo que impide su utilidad comparativa.

En el más reciente CORPES las cifras si están normalizadas. Por ejemplo, la Figura 1.22 muestra los resultados de frecuencia para la palabra *coger* en diferentes dialectos (izquierda) y a través del tiempo (derecha).

Donde el CORPES tiene grandes fortalezas es en mostrar variación por género, dominio y tipo textual. Por ejemplo, la Figura 1.23 muestra la frecuencia para la palabra *lindo* por [tema] (izquierda) y por [tipología] (derecha).

Desafortunadamente, esta categorización textual no es completa. Tal como se dice en el sitio web del CORPES: "la tipología textual se ha incorporado solo a una pequeña parte de los documentos". Por lo tanto, es imposible saber si el 1%, 5% o el 10% de los textos han sido categorizados. Si bien es posible investigar palabras individuales en el CORPES o en otros

Estadísticas

Año	%	Casos
2002	11.34	102
1997	8.00	72
2000	7.23	65
1996	6.45	58
1986	6.22	56
1981	6.00	54
2001	5.78	52
1980	5.33	48
1995	5.33	48
Otros	38.26	344

País	%	Casos
ARGENTINA	28.80	365
ESPAÑA	17.75	225
VENEZUELA	9.31	118
PERÚ	8.68	110
MÉXICO	7.57	96
CHILE	7.02	89
COLOMBIA	4.49	57
URUGUAY	4.49	57
PARAGUAY	2.84	36
Otros	8.99	114

Tema	%	Casos
7.- Ficción.	45.38	600
5.- Ocio, vida cotidiana.	15.73	208
9.- Oral.	13.61	180
4.- Artes.	9.68	128
2.- Ciencias sociales, creencias y pensamiento.	8.85	117
3.- Política, economía, comercio y finanzas.	3.47	46
8.- Miscelánea.	1.51	20
6.- Salud.	0.90	12
1.- Ciencia y Tecnología.	0.83	11

Figura 1.21 Distribución de *lindo* por fecha, país y género en el CREA.

País	Freq	Fnorm.
España	15.118	167,64
Colombia	1.586	73,76
México	1.016	31,43
Cuba	678	68,29
Perú	626	68,31
Argentina	564	22,34
Chile	449	26,64
Venezuela	300	22,68
Puerto Rico	260	67,10
República Dominicana	235	38,17

Período	Freq	Fnorm.
2001-2005	154	1,58
2006-2010	111	1,05
2011-2015	35	0,49

Figura 1.22 Distribución de *lindo* por país y fecha en el CORPES.

Tema	Freq	Fnorm.
Teatro	984	110,92
Relato	514	41,23
Guion	36	37,66
Novela	1.883	33,62
Artes, cultura y espectáculos	386	13,39
Actualidad, ocio y vida cotidiana	440	12,57
Ciencias sociales, creencias y pensamiento	339	9,57
Ciencias y tecnología	76	2,66
Política, economía y justicia	115	2,19
Salud	36	1,92

Tipología	Freq	Fnorm.
Magacines y variedades	11	49,05
Varios	22	47,38
Ficción	3.416	43,72
Entrevista digital	3	31,95
Retransmisiones deportivas	2	31,23
Biografía memoria	43	27,89
Entrevista	87	18,49
Carta al director	1	16,91
Crítica	26	15,05
Opinión	35	10,27

Figura 1.23 Distribución de *lindo* por tema y tipología en el CORPES.

corpus de la RAE, no es posible encontrar todas las palabras o frases cuya frecuencia es más alta en un género, dialecto, o período histórico que en otro (como en las Figuras 1.17, 1.19 y 1.20 de más arriba de el Corpus del Español). También es difícil (y en algunos casos imposible) investigar por cadenas de palabras, tales como la construcción *para* PRON-SUBJ VERB (*para ella entender*), la cual se discutió arriba en la Sección 3.3.

Aunque 2.127.738 textos en el Corpus del Español no han sido categorizados por tópico, es fácil y rápido crear un "Corpus Virtual" para cualquier tópico de interés. Los usuarios pueden crear un Corpus Virtual basándose en los metadatos para los textos (e.g., el título de la página web) o pueden crear un Corpus Virtual basado en palabras de las mismas páginas web. Por ejemplo, en menos de dos segundos, un investigador puede encontrar las páginas web que usan la palabra *inversiones* la mayoría ya sea una frecuencia bruta o una frecuencia normalizada por 1.000 palabras y luego guardar este grupo de textos como un Corpus Virtual. Ellos pueden entonces limitar las búsquedas futuras al Corpus Virtual, comparar entre diferentes Corpus Virtuales o incluso generar una lista de palabras clave desde el Corpus Virtual, como se ve en la Figura 1.24 (palabra clave en el corpus: *inversiones*).

El corpus esTenTen18 del Sketch Engine también permite crear Corpus Virtuales pero no al nivel de el Corpus del Español.

6. Corpus paralelos

El foco de este capítulo ha estado solo en los corpus del español y los corpus paralelos serán objeto de otro capítulo (ver en este volumen el capítulo de Carrió-Pastor y Alonso-Almeida). Los corpus paralelos son, a menudo, no considerados por quienes están interesados en los grandes corpus, pero ellos pueden ser valiosos en sí mismos. Muchos de ellos se basan en documentos gubernamentales, tales como documentos de las Naciones Unidas. Mientras estos corpus pueden ser muy grandes, una importante limitación es que estos textos típicamente representan un lenguaje muy formal en un dominio muy restringido tal como relaciones internacionales o negociaciones de comercio.

Probablemente, más interesantes resultan los grandes corpus paralelos que se basan en programas de televisión o en películas donde el lenguaje es más informal. El sitio web Open Subtitles (www.opensubtitles.org) contiene cientos de millones de palabras de este tipo y esta información ha sido convertida a un formato más accesible por medio de OPUS (http://opus.nlpl.eu/OpenSubtitles-v2018.php). Un ejemplo de ello es el siguiente, donde cada oración en el corpus se alinea con la oración en un segundo idioma.

(src)="73"> They found blood.
(trg)="72"> Encontraron sangre.
(src)="74"> Matches hers.
(trg)="73"> Concuerda con la de ella.
(src)="75"> You see this stuff every day, you know, people getting hurt and killed.

AYUDA	PALABRA (HACER CLIC)	FREC	# TEXTOS	ESPECÍFICO FREC 90 9 TEXTOS	CORPUS ENTERO	ESPERADO
1	REGALÍA	238	11	62.9	794	3.8
2	INVERSIONISTA	986	31	26.6	7,761	37.0
3	BILLÓN	404	13	23.0	3,681	17.6
4	HIDROCARBURO	125	12	19.6	1,338	6.4
5	DIVIDENDO	256	12	18.2	2,948	14.1
6	LINEAMIENTO	91	9	18.1	1,054	5.0
7	INVERSOR	559	21	18.1	6,480	30.9
8	ATRIBUCIÓN	526	9	17.3	6,366	30.4
9	INCISO	939	12	15.9	12,355	58.9
10	DÓLAR	2569	40	14.4	37,432	178.5

Figura 1.24 Palabras clave en el Corpus Virtual *[inversiones]* (Corpus del Español).

(trg)="74"> Ves diariamente estas cosas . . .
(trg)="75"> Gente siendo herida y asesinada . . . y . . . no lo sé.
(src)="76"> And I don't.
(src)="77"> It freaks me out.
(trg)="76"> Me vuelve loco.
(src)="78"> Just the thought of that happening to you, you know . . .
(trg)="77"> El solo pensar que te suceda algo, ya sabes . . .

Esta información ha sido empleada por sitios web como reverso.net y linguee.com para crear corpus muy grandes y poderosos. Los corpus paralelos son probablemente muy utilizados para realizar traducciones, mientras que los otros corpus descritos más arriba se emplean principalmente para investigación.

7. Conclusiones y direcciones futuras

En los apartados precedentes, nos hemos referido a tres colecciones diferentes de corpus que proveen grandes corpus del español (más grandes que un billón de palabras): Sketch Engine, Corpus de la Web y Corpus del Español. Además, los corpus paralelos brindan acceso a información multilingüe de cientos de millones de palabras de español informal (televisión y películas). Corpus más pequeños con 100 a 300 millones de palabras también están disponibles desde la RAE.

Los corpus proveen un alto nivel de funcionalidad en términos de búsquedas de frecuencias de palabras, de patrones (por medio de construcciones gramaticales y n-gramas), significados léxicos y usos (por medio de colocaciones), y la habilidad para examinar la variación a través del tiempo y entre géneros y dialectos. Por todo ello no sorprende que la información alcanzada a partir de estos grandes corpus sea mucho más rica y brinde muchos más aprendizajes que a partir de corpus más pequeños. Cada uno de los corpus que hemos discutido tiene sus fortalezas y debilidades y creemos que lo mejor es considerar a cada uno de estos corpus como una "caja de herramientas", a partir de la cual cada investigador empleará las herramientas que le resulten más útiles para la tarea encomendada, más que concentrarse en un solo corpus particular para sus necesidades de investigación.

Tal como hemos visto a lo largo de este capítulo, en la actualidad, la construcción de grandes corpus textuales crecientes no representa muchas dificultades. Existen amplios desarrollos tecnológicos que permiten contar con capacidades instaladas en términos de equipamiento para almacenar ingentes cantidades de información, así como poderosos programas computacionales de gran versatilidad para ejecutar una amplia variedad de tareas. En este sentido, el futuro es promisorio y los retos no parecen tener límites. Sin embargo, un desafío permanente de mayor complejidad, tal como hemos visto previamente, está en los requerimientos del marcaje morfosintáctico y de otras categorías textuales y discursivas. La revisión manual humana en muchos casos sigue siendo un factor fundamental. No disponemos hasta hoy de un sistema automático confiable de este tipo que funcione completamente independiente en su totalidad y en el cual podamos confiar completamente. Junto a ello, tampoco hemos logrado aún desarrollar sistemas sofisticados que permitan identificar y clasificar rasgos que involucren otros dominios o dimensiones constitutivas de los textos más allá de las palabras como, por ejemplo, los artefactos multimodales de tipo tabla, gráfico, esquema, mapa o diagrama. Ni mucho menos hemos logrado avanzar en desarrollar sistemas computacionales que logren identificar y catalogar las relaciones semánticas de dependencia que vinculan, por ejemplo, un segmento textual de palabras con un gráfico o con una ilustración, ambos como partes constitutivas de

un texto (Parodi 2010 a y b). Estos, entre otros desafíos, impelen a la lingüística de corpus a desafiarse a sí misma y proyectar los nichos de investigación a futuro.

8. Lecturas adicionales

Davies, M. 2019. "The Best of Both Worlds: Multi-billion Word 'Dynamic' Corpora". En *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019*, eds. P. Bański, A. Barbaresi, H. Biber y E. Breiteneder, 23–28. Mannheim: Leibniz-Institut fur Deutsche Sprache.

Schäfer, R. y F. Bildhauer. 2012. "Building Large Corpora from the Web Using a New Efficient Tool Chain". En *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12), eds. N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk y S. Piperidis, 486–493. Estambul: ELRA.

Referencias citadas

Bick, E. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.

Davies, M. 2008. "Spanish and Portuguese Corpus Linguistics". *Studies in Hispanic and Lusophone Linguistics* 1: 149–186.

Davies, M. 2010. "Creating Useful Historical Corpora: A Comparison of CORDE, the Corpus del Español, and the Corpus do Português". En *Diacronía de las Lenguas Iberorromances: Nuevas Perspectivas desde la Lingüística de Corpus*, ed. A. Enrique-Arias, 137–166. Frankfurt/Madrid: Vervuert/Iberoamericana.

Davies, M. 2015. "Corpora: An Introduction". En *Cambridge Handbook of English Corpus Linguistics*, eds. D. Biber y R. Reppen, 11–31. Cambridge: Cambridge University Press.

Davies, M. 2017. "Using Large Online Corpora to Examine Lexical, Semantic, and Cultural Variation in Different Dialects and Time Periods". En *Corpus-Based Sociolinguistics*, ed. E. Friginal, 19–82. Londres: Routledge.

Davies, M. 2018a. "Corpus-based Studies of Lexical and Semantic Variation: The Importance of both Corpus Size and Corpus Design". En *From Data to Evidence in English Language Research*, eds. S. Carla, T. Nevalainen y I. Taavitsainen, 34–55. Leiden: Brill.

Davies, M. 2018b. "Uso del Corpus del Español y los corpus relacionados para la lexicografía histórica española". En *Historia del Léxico Español y Humanidades Digitales*, eds. D. Corbella, A. Fajardo y J. Lagenbacher-Liebgott, 49–76. Berlín: Peter Lang.

Davies, M. y K. Davies. 2017. *A Frequency Dictionary of Spanish: Core Vocabulary for Learners*. Londres: Routledge.

Davies, M. y R. Fuchs. 2015. *Expanding Horizons in the Study of World Englishes with the 1.9 Billion Word Global Web-Based English Corpus (GloWbE)*. Londres: John Benjamins.

Firth, J. R. 1957. *Papers in Linguistics 1934–1951*. Londres: Oxford University Press.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J, Michelfeit, P. Rychlý y V. Suchomel. 2014. "The Sketch Engine: Ten Years on". *Lexicography* 1: 7–36.

Parodi, G. 2010a. "Research Challenges for Corpus Cross-linguistics and Multimodal Texts". *Information Design Journal* 18 (1): 69–73.

Parodi, G. 2010b. "Multisemiosis y lingüística de corpus: artefactos (multi)semióticos en los textos de seis disciplinas en el Corpus PUCV-2010". *Revista de Lingüística Teórica y Aplicada (RLA)* 48 (2): 33–70.