

USO DE CORPUS EN CLASE DE ELE

La lengua real como modelo



difusión



Los corpus lingüísticos se han mostrado en las últimas décadas dentro de la lingüística como una línea de investigación con un marcado carácter transversal que está generando propuestas notables. En este sentido, hay que señalar un hecho claramente relevante: la investigación basada en estas herramientas ha adquirido tanta presencia en los últimos tiempos que ha dado el salto desde ser considerada un simple instrumento hasta alzarse como eje central de su propio campo de investigación. Así, hemos pasado de hablar de corpus para la lingüística a delimitar y desarrollar una lingüística de corpus.

Es evidente que la particularidad principal de los corpus como interfaces de investigación reside en su naturaleza tecnológica. Como instrumento de trabajo va más allá de los útiles tradicionales de la investigación lingüística y abre un campo de posibilidades impensable hace solo pocas generaciones. De alguna manera, el concepto de corpus se ha hecho equiparable, como forma de acceso al conjunto infinito de producciones de la lengua, a otros instrumentos más prestigiosos y con mayor presencia en nuestro día a día, como el diccionario o la gramática.

NEUS SANS Y FRANCISCO HERRERA

En la mayoría de los programas de estudios no se incluye el uso de corpus, pero, paradójicamente, muchos alumnos los emplean para mejorar y corregir sus producciones de texto, porque acceder a ellos es tan fácil como realizar una búsqueda en Google.

Pero ¿para qué puede usarlos el docente? Durante años, en clase hemos enseñado lo que creíamos más habitual desde la introspección, pero lo cierto es que a veces las formas que se nos ocurren como más prototípicas no son las más habituales en la lengua. Pongamos que quiero expresar una opinión; una de las primeras opciones que se me ocurrirá es “Yo creo que...”; y sin embargo una forma mucho más habitual de hacerlo es comenzar una frase con “Pues yo...”. La única manera de saber cuáles son esas estructuras más usadas (y por lo tanto las que querremos llevar a clase) es contar con muestras de habla real. Y eso es lo que encontramos en los corpus: muestras que nos sirven para enseñar la lengua real y no una lengua creada en un despacho en que la gramática es perfecta y nunca se dejan frases suspendidas.

WENDY ELVIRA-GARCÍA

**PRÓLOGO · PREFACIO 1 LINGÜÍSTICA DE CORPUS Y EL
PENSAMIENTO LINGÜÍSTICO: EL EMPIRISMO 2 ¿QUÉ ES LA
LINGÜÍSTICA DE CORPUS? 3 CÓMO HABLAR DE CORPUS:
CONCEPTOS CLAVE Y TERMINOLOGÍA EN LINGÜÍSTICA DE
CORPUS 4 TIPOLOGÍA Y DISEÑO DE CORPUS 5 APLICACIONES
DE LOS CORPUS EN LA ACTUALIDAD 6 CORPUS PARA
INVESTIGAR SOBRE EL ESPAÑOL LENGUA EXTRANJERA 7 LOS
CORPUS EN LA CREACIÓN DE MATERIALES 8 ACTIVIDADES
CON CORPUS PARA LLEVAR AL AULA 9 LOS CORPUS PARA
LA CORRECCIÓN 10 CORPUS PARA EL APRENDIZAJE
AUTÓNOMO DEL ALUMNO 11 SECUENCIAS DIDÁCTICAS DE
CORPUS Y PROGRESIÓN DEL USO DE CORPUS EN UN CURSO**

difusión



DL: B 6947-2021

ISBN: 978-84-18625-32-9



9 788418 625329

cuadernos de didáctica

USO DE CORPUS EN CLASE DE ELE

La lengua real como modelo

WENDY ELVIRA-GARCÍA
UNED



CUADERNOS DE DIDÁCTICA

Colección dirigida por Francisco Herrera y Neus Sans

USO DE CORPUS EN CLASE DE ELE

La lengua real como modelo

AUTORA: Wendy Elvira-García

EDICIÓN: Francisco Herrera y Neus Sans

REDACCIÓN: Roberto Castón (ilusionoptica.es)

CORRECCIÓN ORTOGRAFÍA: Marina López

DISEÑO DE CUBIERTA E INTERIORES: Laurianne López Barrera

MAQUETACIÓN: Aleix Tormo

ILUSTRACIÓN: Laurianne López Barrera

© La autora y Difusión S.L. Barcelona 2021

978-84-18625-32-9

Impreso en la UE

Queda prohibida cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con la autorización de los titulares de la propiedad intelectual. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (art. 270 y ss. Código Penal).



C/ Trafalgar, 10, entlo. 1ª
08010 Barcelona - España
Tel.: (+34) 932 680 300
Fax: (+34) 933 103 340
editorial@difusion.com

www.difusion.com

ÍNDICE

6	Prólogo
10	Prefacio
16	1 / Lingüística de corpus y el pensamiento lingüístico: el empirismo
24	2 / ¿Qué es la lingüística de corpus?
36	3 / Cómo hablar de corpus: conceptos clave y terminología en lingüística de corpus
46	4 / Tipología y diseño de corpus
62	5 / Aplicaciones de los corpus en la actualidad
68	6 / Corpus para investigar sobre el español lengua extranjera
78	7 / Los corpus en la creación de materiales
88	8 / Actividades con corpus para llevar al aula
104	9 / Los corpus para la corrección
110	10 / Corpus para el aprendizaje autónomo del alumno
114	11 / Secuencias didácticas de corpus y progresión del uso de corpus en un curso
122	Bibliografía
130	Solucionario
146	Glosario

PRÓLOGO

Como editores de este volumen de *Cuadernos de didáctica*, la primera entrega de la colección que cuenta con una sola autora y con la perspectiva ceñida a un tema tan específico, lo primero que se nos plantea es la necesidad real de este prólogo. Por supuesto, el libro se defiende solo ante los lectores, que ya desde el índice podrán medir la claridad del planteamiento y la relevancia de su publicación. Sin embargo, nos gustaría subrayar algunos aspectos que son realmente significativos en esta publicación y que nos han llevado a proponer su aparición dentro de esta serie didáctica.

En primer lugar, hay que remarcar el alcance de su objeto de estudio. Los corpus lingüísticos se han mostrado en las últimas décadas dentro de la lingüística como una línea de investigación con un marcado carácter transversal que está generando propuestas notables. En este sentido, hay que señalar un hecho claramente relevante: la investigación basada en estas herramientas ha adquirido tanta presencia en los últimos tiempos que ha dado el salto desde ser considerada un simple instrumento hasta alzarse como eje central de su propio campo de investigación. Así, hemos pasado de hablar de corpus para la lingüística a delimitar y desarrollar una lingüística de corpus.

Es evidente que la particularidad principal de los corpus como interfaces de investigación reside en su naturaleza tecnológica. Como instrumento de trabajo va más allá de los útiles tradicionales de la investigación lingüística y abre un campo de posibilidades impensable hace solo pocas generaciones. De alguna manera, el concepto de corpus se ha hecho equiparable, como forma de acceso al conjunto

infinito de producciones de la lengua, a otros instrumentos más prestigiosos y con mayor presencia en nuestro día a día, como el diccionario o la gramática.

Esa tecnología invisible que son los corpus mantiene, por lo tanto, una relación increíblemente fructífera con la lingüística y con toda probabilidad todavía nos quedan por descubrir un gran conjunto de usos y aplicaciones novedosos que nos permitirán entender la naturaleza maleable y exitosa de estos enfoques.

Sin embargo, si analizamos ese mismo tipo de vínculos con la didáctica y en concreto con la enseñanza de segundas lenguas, se hace evidente que queda mucho trayecto todavía por recorrer. Esa invisibilidad que mencionábamos antes se hace mucho más patente cuando hablamos de iniciativas para llevar los corpus a los procesos de enseñanza y aprendizaje de idiomas, aunque, poco a poco, el camino se va desbrozando con propuestas como las que tenemos ahora mismo entre manos.

En este sentido, consideramos que la autora ha pulsado todas las cuerdas necesarias para que su propuesta nos haga reflexionar y actuar sobre esta tendencia y nos permita sacar a los corpus de la zona menos visible para darles la relevancia que merecen también como instrumentos didácticos. Sin duda, el hecho de que el libro cuente con una estupenda batería de actividades nos va a permitir entender mejor esa doble naturaleza investigadora y divulgadora.

Queremos creer, por lo tanto, que los corpus van a alcanzar por fin ese hueco que se merecen en la atención de los docentes de español, así que no nos queda más que agradecer a la autora por todo el esfuerzo que ha invertido para que así sea.

Francisco Herrera y Neus Sans

PREFACIO

En el mundo anglosajón, editoriales de manuales de inglés para extranjeros de prestigio, como Cambridge University Press, llevan décadas presumiendo de que sus manuales son los mejores porque están basados en corpus y eso los hace precisos y actuales. Sin embargo, las editoriales de español para extranjeros han tardado mucho más en incluir material de corpus en sus manuales y aún hoy su uso no está generalizado. Por no hablar de su uso en clase, y es que en la mayoría de los programas de estudios no se incluye el uso de corpus, pero, paradójicamente, muchos alumnos los emplean para mejorar y corregir sus producciones de texto, porque acceder a ellos es tan fácil como realizar una búsqueda en Google.

Pero ¿para qué puede usarlos el docente? Durante años, en clase hemos enseñado lo que creíamos más habitual desde la introspección, pero lo cierto es que a veces las formas que se nos ocurren como más prototípicas no son las más habituales en la lengua. Pongamos que quiero expresar una opinión; una de las primeras opciones que se me ocurrirá es “Yo creo que...”; y sin embargo una forma mucho más habitual de hacerlo es comenzar una frase con “Pues yo...”. La única manera de saber cuáles son esas estructuras más usadas (y por lo tanto las que queremos llevar a clase) es contar con muestras de habla real. Y eso es lo que encontramos en los corpus: muestras que nos sirven para enseñar la lengua real y no una lengua creada en un despacho en que la gramática es perfecta y nunca se dejan frases suspendidas.

Este libro es una guía y un acicate para que los profesores de español lengua extranjera incluyan muestras de corpus en su práctica docente. En él, hablaré de cómo usar corpus para crear manuales y clases, pero también de cómo

podemos enseñar a nuestros alumnos a usarlos para que sean ellos mismos los que investiguen la lengua y descubran las reglas del español a partir del estudio de casos.

La idea de este manual nació a partir de la docencia de la asignatura “Lingüística de corpus y enseñanza del español como segunda lengua” que se ofrece como optativa en el máster en Formación de profesores de español como segunda lengua de la Universidad Nacional de Educación a Distancia. Preparando la asignatura, me encontré con que no existían manuales (ni prácticamente más información que algunos *webinars* y *podcasts*) sobre cómo llevar los corpus a clase. No tenía un libro guía para la asignatura que cubriera la parte que a mí me parecía más interesante: el uso que puede hacer el profesorado de los corpus. Creo que este manual puede ser de utilidad para cualquier profesor o formador que se plantee el uso de corpus en el aula de español, pero también para otros profesionales que se quieran acercar al uso del corpus para la investigación.

Y es que presenta una diferencia esencial con los (pocos) materiales publicados para el uso de corpus y la enseñanza del español lengua extranjera. Normalmente, los materiales para el trabajo de corpus y ELE explican qué son los corpus, cómo se usan (lo que recoge la primera parte de este manual) y dan una panorámica de los corpus que hay disponibles, pero dan pocas o ninguna idea de qué tipo de actividades se pueden realizar en clase. Este manual explica qué es un corpus y cómo usarlo, pero, además, da una explicación detallada y ejemplificada de en qué campos se puede usar corpus; como creación de materiales y ejemplos basados en corpus, pero también actividades de clase. Para ello, se proponen actividades concretas y se aborda cómo realizar secuencias didácticas con corpus. Es decir, se trata los corpus como una herramienta a la disposición del profesor, pero también de los estudiantes a partir de actividades donde son los alumnos los que buscan en un corpus para, por ejemplo, extraer una regla gramatical de manera inductiva.

En el texto se da por hecho que el lector está familiarizado con el mundo del ELE y que las metodologías y términos propios del área no le son desconocidos. Términos como *interlengua*, *aprendizaje por tareas*, *énfoque léxico*, *comunicativo*, *secuencia didáctica*, *aprendizaje inductivo* o *clase invertida* se usarán con una introducción muy breve. Por ello, es recomendable consultar un manual general sobre la práctica docente de español como lengua extranjera, en el caso de que no se tengan esos conocimientos. Pueden ser recomendables manuales como el de Andión Herrero, González Sánchez, & San Mateo Valdehita (2019) o simplemente acudir al Diccionario de términos clave de ELE (Varios Autores, 2008), disponible en línea.

Por último, introduzco aquí cómo está organizado este manual. Se organiza en 11 temas, que pertenecen a dos bloques temáticos: el primero trata de la lingüística de corpus en general y el segundo de el uso de los corpus en ELE. A estos dos bloques les siguen la bibliografía y el solucionario de las actividades propuestas.

El primer bloque consta de cinco temas y en él se introducen los conceptos básicos de la lingüística de corpus, necesarios para poder hacer búsquedas y explotar corpus. Pero no, el manual no contiene un análisis pormenorizado de cómo anotar y estándares de anotación, niveles de etiquetaje o el uso de corpus para entrenamiento de sistemas de procesamiento del lenguaje natural, que sí se incluirían en un curso de corpus más enfocado a la lingüística computacional. Este manual se centra en los corpus como una herramienta para el uso en el aula de español. Por eso, pese a que este primer bloque trata conceptos generales que se pueden encontrar en un manual de lingüística de corpus, se intenta mantener siempre una perspectiva de profesor de español ahondando en los detalles de diseño y terminológicos que pueden ser útiles al docente y obviando otros. Además, siempre que ha sido posible, se ha intentado llevar los ejemplos al terreno del español segunda lengua.

El segundo bloque contiene seis temas y en ellos se ahonda en cada una de las tareas que se pueden realizar en clase o preparar con un corpus de ELE. Comienza con la aplicación de los corpus a la investigación. En esa parte, se tratan los corpus de aprendientes, los únicos específicos del mundo de la enseñanza de segundas lenguas, que nos sirven para descubrir las dificultades de nuestros alumnos y para realizar trabajos de investigación (como trabajos de final de máster o tesis doctorales). Tras ese tema, se pasa a tratar de lleno la labor del profesor y se trabaja el uso de corpus en la creación de materiales, ya sean manuales de ELE al uso (para aquellos que trabajan como editores) o las fotocopias de clase; se trata también cómo llevar los corpus al aula para que los puedan usar los estudiantes, ya sea para explicar gramática, léxico, cultura o pronunciación; cómo usarlos en nuestras correcciones en clase y para que los alumnos se puedan autocorregir; y cómo los estudiantes pueden usar los corpus para crear redacciones más correctas y ricas. Por último, se da una visión integradora de todo ello a partir de un capítulo dedicado a la creación de secuencias didácticas con corpus.

Independientemente del bloque al que pertenezca, cada tema consta de diferentes epígrafes de teoría, una selección de lecturas para profundizar en el temario (capítulos de libros o artículos) y una serie de actividades. Las actividades forman parte del temario en el sentido en que, para saber usar corpus en clase de ELE, es

necesario conocerlos y, para conocerlos, es necesario haber usado sus interfaces (páginas web) y visto sus posibilidades. Además, en la segunda parte, las actividades son muchas veces ejemplos de cómo llevar los corpus a las aulas.

En la parte teórica del libro no se explican las posibilidades de cada corpus. Si un corpus está lematizado o no, o el tipo de búsquedas que se pueden realizar es algo que se debe consultar en la descripción de cada corpus, por lo que resultaría redundante incluir un resumen aquí. Por lo tanto, será necesario aprender a encontrar esa información tanto en la interfaz de cada corpus como a partir de las búsquedas, es decir, es a partir de la explotación guiada que se hará de los corpus en las actividades. Así, a partir de su uso, aprenderemos y reflexionaremos sobre sus limitaciones. Por eso, es especialmente importante seguir los ejercicios tanto en el primer bloque como en el segundo, porque solo a partir del conocimiento generado a partir de la práctica, se podrán aplicar las búsquedas a la creación de materiales de ELE, la creación de una secuencia didáctica o la corrección. En las actividades de los temas, también se da una lista de los términos importantes, que, si se van definiendo en cada uno de ellos, llevarán a la creación de un glosario. Como he dicho, todas las actividades y las palabras propuestas para su definición tienen un solucionario al final del libro.

Y, sin detenerme más, pasamos al contenido. Espero que este manual pueda aportarte algunas ideas útiles y atractivas para llevar a clase o para tus investigaciones y también espero que, después de seguir el libro, me ayudes en la labor de librar a los corpus de su mala fama en las aulas.

Wendy Elvira-García

1

LINGÜÍSTICA DE CORPUS Y EL PENSAMIENTO LINGÜÍSTICO: EL EMPIRISMO

1. LINGÜÍSTICA DE CORPUS Y EL PENSAMIENTO LINGÜÍSTICO: EL EMPIRISMO

El español lengua extranjera (de ahora en adelante, ELE), igual que la lingüística general, se ha visto siempre fuertemente influido por las corrientes de pensamiento de su época. Así, en una época en que la lingüística se entendía como el estudio de la gramática y especialmente de la traducción de lenguas clásicas, en enseñanza de lenguas extranjeras el método de la gramática-traducción vivía sus mejores tiempos (Melero, 2000). Fue solo a partir del funcionalismo de Halliday que se empezaron a poner de moda métodos centrados en las funciones comunicativas del lenguaje (Brumfit, 1984; Halliday, 1985). Esto es importante porque la lingüística de corpus es también hija de esos mismos cambios de pensamiento y, en consecuencia, de paradigma y de filosofía. Por ello, antes de entrar de lleno en la lingüística de corpus, vamos a intentar explicar las razones que llevaron a su nacimiento.

La lingüística tradicional basaba sus estudios en la figura del gramático sabio que, detrás de las torres de libros de su mesa de despacho, pensaba en usos gramaticales y agramaticales (no válidos) de la lengua y, así, llegaba a la conclusión, por ejemplo, de que en español (1)

(1)

Creo que + VERBO INDICATIVO

No creo que + VERBO SUBJUNTIVO

Fillmore llamaba a este tipo de lingüista: “lingüista de sillón orejero” (Fillmore, 1991). En clase, no suele haber sillones orejeros, y, sin embargo, en el momento en el que escribimos en la pizarra (2)

(2)

—¿Qué hace Juan?

—Juan come manzanas.

no hay mucha diferencia entre el gramático sabio y nosotros. Nos hemos inventado el ejemplo tal y como hacía el gramático sabio. Pero ¿cuántas veces en su vida un alumno va a usar la frase “Juan come manzanas”?, un enunciado informativo

neutro con sujeto explícito (llamado *Juan*). En español, eso no es habitual. Cuando queremos informar sobre algo, omitimos el sujeto (excepto cuando hay un contraste informativo). Por ello, el ejemplo de (3) sería algo más realista

(3)

—¿Qué hace Juan?

—Come manzanas.

Sin embargo, todos hemos estado con la tiza en alto en un momento de tensión ante la clase en el que queremos encontrar un ejemplo explicativo, prototípico, de uso frecuente, léxico nivel básico, y lo único que te viene a la cabeza es el pobre Juan.

La alternativa a los ejemplos inventados, ya sea en gramática teórica o aplicada al ELE, pasa por el uso de ejemplos de lengua real. Es decir, por la disponibilidad de muestras de habla en su contexto. Todo cambio de paradigma en investigación necesita un entorno propicio y la comprensión de la necesidad de estas “muestras de habla en su contexto”, es decir, de *datos*, fue el primer paso para la creación de la lingüística de corpus.

1.1. EMPIRISMO CONTRA RACIONALISMO: DOS MANERAS DE ENTENDER LA LINGÜÍSTICA

A principios del siglo xx, los lingüistas empezaron a interesarse por una disciplina nueva muy ligada a la sociología: la dialectología (Alvar, 1969). Sumergidos en nuevas tendencias de la lingüística (introducidas por Saussure) que separaban la *lengua* como ente abstracto e inaprensible y el *habla*, como sus realizaciones por parte de los hablantes, descubrieron que solo el habla podía ser un objeto de estudio. Por ello, los gramáticos sabios empezaron a salir de sus despachos, donde inventaban ejemplos que sirvieran a sus teorías sobre la lengua, y salieron a la calle a escuchar y a anotar cómo hablaba la gente. Para ello, se empezaron a realizar encuestas a los hablantes en las que se preguntaba cómo se llamaba cierto objeto en diferentes puntos del mapa y de esta manera se obtuvieron datos reales sobre las diferentes denominaciones que recibía cada referente. Esos datos, relacionados con el punto de encuesta donde se habían documentado, se convertían después en grandes atlas lingüísticos.

En paralelo, lingüistas estadounidenses, como Boas o Sapir, empezaron a interesarse por las lenguas nativas de sus territorios dando el pistoletazo de salida a lo que conocemos como antropología lingüística (Koerner, 2003). Se trataba de lenguas que nunca habían sido estudiadas desde un punto de vista lingüístico-gramatical y para las que, además, los lingüistas no eran hablantes nativos. Esto

hizo que fuera imprescindible realizar lexicones (pequeños diccionarios) y tomar nota de las frases que oían (otra vez, datos).

Con todo ello, se empezó a aplicar el trabajo de campo de manera general en los departamentos de lingüística. Los gramáticos estructuralistas (discípulos de Saussure) empezaron a ver las ventajas de contar con muestras de habla real para poder explicar los fenómenos de la lengua a la vez que constataban que la gente de la calle no hablaba como en sus gramáticas. En definitiva, se volvieron defensores del empirismo, de aportar pruebas para sus teorías.

Contar con datos auténticos, evitar la influencia que pueda tener la propia variedad (dialecto o acento) del hablante, contar con las variedades de más gente y, por lo tanto, con más diversidad y cantidad de muestras, y poder cuantificar los datos (por ejemplo, calculando el tanto por ciento de gente que usa *dalle* o *azada* en cada región) eran algunas de las ventajas que los lingüistas obtenían al salir de su despacho y preguntar a la gente cómo hablaba.

Pero, cuando parecía que la tendencia se iba a asentar y que el uso de datos reales se convertiría en norma, llegó Noam Chomsky y se convirtió en la moda. Chomsky es un lingüista que venía de un bagaje matemático y, por tanto, estaba más interesado en formalizar el lenguaje que en la lengua en sí. Además, seguía una corriente de pensamiento más racionalista. Creía que contar con datos reales solo servía para hacer imposible extraer generalizaciones. En sus propias palabras:

[m]y judgment, if you like, is that we learn more about language by following the standard method of the sciences. The standard method of the sciences is not to accumulate huge masses of unanalyzed data and to try to draw some generalization from them (Chomsky, 2004) citado en (Taylor, 2008).

Y, en parte, tenía razón. El uso de datos reales no estaba exento de problemas. Para empezar, al recopilar datos, se recogen muestras de habla, pero no de la lengua. La lengua, entendida como la facultad del lenguaje, es una abstracción, la capacidad universal que tenemos los humanos de hablar. En este sentido, las muestras de habla nos pueden aportar datos únicamente sobre lo que pasa cuando esa facultad universal se concretiza en una lengua, un dialecto, un hablante... Pero hacer inferencias sobre las características universales, comunes de las lenguas, sigue siendo tarea del lingüista. Los datos nunca van a aportar explicaciones de por qué pasa algo.

Además, los datos nunca son exhaustivos. Es decir, no contienen una variedad al completo, porque para eso se necesitarían todas las muestras de lengua emitidas por todos los hablantes en toda la historia y en todos los registros (oral, escrito, planificado,

conversacional...) y eso es imposible. Por eso, es muy importante que los datos sean representativos (y a este concepto volveremos más adelante).

Por lo tanto, los datos nos mostrarán lo que es cierto en la variedad: por ejemplo, la gente de Madrid dice *coche* y no *auto*, pero que yo no haya encontrado ningún hablante madrileño que diga *auto* no quiere decir necesariamente que esa persona no exista. A eso lo llamamos “no aportar evidencia negativa”. Vamos a ver un ejemplo más. Pongamos que entrevistamos a un grupo de alumnos de Filología y les hacemos contar una historia en la que unos niños están jugando al juego de las sillas. Todos ellos acaban con una frase del tipo “al final, el niño se tiene que quedar de pie porque no *había sillas* libres”. ¿Esto querrá decir que una frase del tipo “no *habían sillas* libres” no se produce nunca en español? No, solo quiere decir que nuestro grupo de alumnos no lo dice, ya que, probablemente, eso se deba a que son estudiantes de Filología que han cursado una asignatura de normativa del español y están hablando con una profesora y no con sus amigos. Por lo tanto, en el mejor de los casos, esta muestra refleja algo que es cierto en nuestro grupo de estudiantes, pero no en la población general.

La última crítica que se hace a los estudios empíricos, es decir, a aquellos estudios que están basados en datos, tiene que ver justamente con este concepto de representatividad. Para que los datos sean representativos de una población más o menos general se necesitan muchos datos. Se necesitan tantos datos, que para un solo humano es imposible leerlos y hacerse una idea de ellos para clasificarlos. Eso hizo que durante la primera mitad del siglo xx la mayor parte de los datos que se recogieron, sobre todo en atlas lingüísticos, estuvieran ahí, disponibles, pero no se podían analizar, porque no existían técnicas para hacerlo (Abercrombie, 1965). Si alguna vez has trabajado con atlas y has visto láminas con diferentes isoglosas para cada palabra y has intentado a partir de esas isoglosas (las cuales acaban cada una en un pueblo diferente) establecer la frontera entre dos dialectos, entenderás cómo esa cantidad de datos se hace muy difícil de gestionar. Esta limitación se superó en el momento en el que los ordenadores tuvieron capacidad suficiente para realizar análisis y hoy en día no supone ningún problema, ya que las nuevas técnicas de análisis de datos (*data analysis*) se alimentan precisamente de una cantidad ingente de datos (*big data*), como por ejemplo sería el *feed* de Twitter, que genera al día 500 millones de tuits, de los cuales, el 4,7% son en español (Hong, Convertino, & Chi, 2011).

En el lado opuesto de la balanza, el racionalismo crea explicaciones de la lengua, teorías que son válidas para toda la lengua o incluso para la facultad del lenguaje (de

hecho, ese era el objetivo de la gramática generativa de Chomsky) y después crea ejemplos para dar soporte a sus teorías. Esos ejemplos son ejemplos limpios, fáciles de entender y bonitos, como “Juan come manzanas”, pero están totalmente sesgados por el investigador. Su idiolecto (sus maneras de decir individuales) se da como la norma imperante sin haber comprobado si realmente son formas válidas para toda la población o si la palabra *cuchufli*, que a él le parecía tan normal, solo se usa en su casa.

Afortunadamente para nosotros, algunos lingüistas seguían necesitando datos para poder realizar sus estudios, como los antropólogos o los dialectólogos. Otros, como los estudiosos de *Second Language Acquisition* (SLA), descubrieron muy pronto que sus técnicas eran mucho más eficaces si analizaban los datos específicos de su población. Es decir, los datos (principalmente los errores) que sus alumnos generaban. Por ejemplo, analizando los errores de sus alumnos podían localizar en qué temas tenían que poner el foco en años venideros e incluso dar con nuevas maneras de enseñarles y también incluir ejemplos de uso real en los materiales (Pitkowski & Gamarra, 2009).

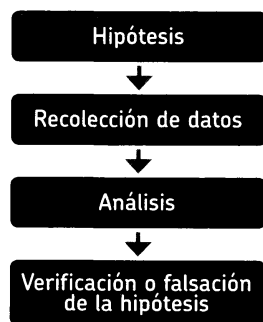
Esto crea una gran separación en lingüística que tiene continuidad hasta nuestros días: lingüistas racionalistas, que usan su intelecto para analizar; y lingüistas empiristas, que se basan en los datos.

1.2. MÉTODO CIENTÍFICO EN LA PRÁCTICA LINGÜÍSTICA

Hoy en día, la mayoría de los lingüistas somos un poco de ambos y es que al seguir el método científico hacemos uso tanto del intelecto como de los datos.

Neither the corpus linguist of the 1950s, who **rejected intuition**, nor the general linguist of the 1960s, who **rejected corpus data**, was able to achieve the **interaction of the data coverage and the insight** that characterize the many successful corpus analyses of recent years. Geoffrey Leech (1991)

En cualquier trabajo de investigación actual, se puede encontrar una hipótesis. Por ejemplo, “los alumnos que cuya lengua materna no tenga vibrante alveolar múltiple [r] tendrán más problemas con ella”. Esta hipótesis en parte está basada en nuestra razón; nosotros sabemos que esto es así, pero también puede estar basada en nuestra experiencia, en datos. Una vez que hemos determinado la hipótesis, recogeremos datos para verificarla o falsarla y, una vez tengamos el análisis de esos datos (probablemente con el porcentaje de alumnos que tienen problemas con [r] clasificados por su lengua materna), podremos concluir si nuestra hipótesis era cierta o no (esquema 1).



Esquema 1. Proceso del método científico

Por lo tanto, aunque el racionalismo sigue formando parte de la mayoría de las investigaciones, hoy en día no se concibe un análisis lingüístico en el cual no haya datos. El método científico, compartido por todas las disciplinas del conocimiento, exige que se ofrezcan pruebas de los hallazgos y esas pruebas son los datos, datos que en lingüística llamamos *corpus*.

1.3. LA LLEGADA DE UN NUEVO SIGLO: ESTUDIOS CORPUS-DRIVEN, APRENDIZAJE DATA-DRIVEN

Elena Tognini-Bonelli dio nombre a una partición que ha sido muy útil en los estudios de corpus desde entonces: estudios basados en corpus (*corpus-based*) y orientados al corpus (*corpus-driven*) (Tognini-Bonelli, 2001). Los primeros estudios eran los que se venían realizando hasta la fecha. En ellos, a partir de una hipótesis, se usaba un corpus para validarla. Por ejemplo, ante la hipótesis de que los anglófonos dirán *la gente *son* por influencia de su lengua materna, se plantea una búsqueda en corpus que devuelva las instancias (más tarde, veremos que en corpus se llaman *ocurrencias*) de *gente* seguido del verbo *ser* y el programa devuelve la frecuencia del singular y del plural. El segundo tipo, *corpus-driven*, observa el corpus y busca patrones y regularidades sin apriorismos ni hipótesis y, a partir de esas observaciones, construye una hipótesis explicativa. Si usamos el ejemplo anterior, para llegar a la misma conclusión, se observaría el corpus (por ejemplo, el conjunto de redacciones de los alumnos anglófonos de una clase), se haría una lista de los errores más frecuentes y, quizá, aparecerían cosas como *la gente *son*. A partir de ahí, se buscarían explicaciones plausibles sobre por qué se han producido esos errores, en este caso, transferencia de la L1.

La conclusión ha sido la misma, pero en el primer caso, se parte de nuestro conocimiento previo, mientras que en el segundo no se hace ningún tipo de

apriorismo. La implicación más beneficiosa de aplicar este método es que se eliminan del método los posibles prejuicios del investigador.

En el campo del ELE, el mismo concepto se ha aplicado al uso de herramientas que permiten a los estudiantes inferir cuál es la pieza léxica o la forma gramatical que están buscando a partir de muestras de lengua (Buyse & Verlinde, 2013). Lo llamamos *aprendizaje deductivo*. Si no has oído nunca hablar de él, puedes buscarlo en el *Diccionario de términos CLAVE de ELE* (Varios Autores, 2008), disponible en https://cvc.cervantes.es/Ensenanza/biblioteca_ele/diccio_ele/indice.htm.

Un ejemplo de aprendizaje deductivo sería el que se suele hacer con la alternancia entre indicativo y subjuntivo en oraciones afirmativas y negativas con verbos de pensamiento tipo *Creo que llueve* pero *No creo que llueva*, donde, a partir de varios ejemplos, los alumnos pueden deducir cuándo se usa el indicativo y cuándo el subjuntivo. Más adelante, en el libro, usaremos este mismo ejemplo para ver cómo se puede usar un corpus en clase.

También se puede usar el mismo método para conseguir que los alumnos infieran el significado de una pieza léxica a partir del contexto. En este último caso, el corpus de Linguae ha demostrado que ayuda a mejorar la precisión léxica de los estudiantes (Buyse & Verlinde, 2013). Esto se debe a que, a partir de muestras reales de lengua en contexto, los alumnos no solo aprenden el significado de la palabra, si no también sus usos más comunes y las palabras que lo suelen acompañar. Es, en cierta manera, como si en vez de memorizar la entrada de un diccionario de definiciones, como el *Diccionario de la Real Academia de la Lengua* (Real Academia Española, 2014), estuvieran memorizando la entrada de un diccionario de colocaciones o, más bien, un diccionario combinatorio como el *Redes* (Bosque, 2004) o el *Práctico* (Bosque, 2006).

1.4. PARA MÁS INFORMACIÓN

McEnery, T. & Hardie, A. (2013). The history of corpus linguistics. *The Oxford handbook of the history of linguistics*, 727-745.

1.5. ACTIVIDADES

1.5.1. ¿Qué ramas de la lingüística fueron pioneras en la recopilación de datos?

1.5.2. ¿Cuáles son las ventajas del empirismo? ¿Y del racionalismo? ¿Y sus desventajas?

2

¿QUÉ ES LA LINGÜÍSTICA DE CORPUS?

2. ¿QUÉ ES LA LINGÜÍSTICA DE CORPUS?

En el capítulo anterior se ha hablado de las diferencias de pensamiento entre lingüistas que recogen datos y que no recogen datos, y ello nos lleva al tema central de este manual, los corpus, ya que así es como llamamos a los datos en lingüística. En esta sección se van a definir los conceptos de *corpus* y *lingüística de corpus*.

2.1. ¿QUÉ ES UN CORPUS?

Un *corpus* es un conjunto de muestras de habla (orales o escritas) representativas de una variedad lingüística que se usan para realizar investigaciones lingüísticas o para entrenar un sistema de inteligencia artificial (McEnery & Wilson, 1996; Taylor, 2008; Torruella & Llisterri, 1999). Es decir, un corpus son datos y esto es especialmente importante en el mundo actual en el que los sistemas informáticos de inteligencia artificial se alimentan de datos. Se ahondará en esta idea al hablar de aplicaciones de corpus.

Acerquémonos más al concepto de corpus. Un corpus es un conjunto de muestras de habla **digitalizadas y procesables por un ordenador**, lo que nos facilita hacer búsquedas y ordenar los datos. Estas muestras pueden ser habla **escrita u oral**. En el segundo caso, deben ir acompañadas de sus respectivas transcripciones y de manera ideal esas transcripciones estarán alineadas con el audio, como ocurre, por ejemplo, con los subtítulos de una película en que el sonido y el texto aparecen a la vez. Las muestras son necesariamente **textos originales**, es decir, no creados a propósito para tener ejemplos de habla, sino porciones que se hayan usado con una función comunicativa, ya sea en un libro, en una conversación o en cualquier otro intercambio comunicativo, como un chat o una reseña de Amazon. Pueden ser documentos **completos** o también **fragmentos**. Si tenemos un corpus de libros, podemos usar libros enteros o, pongamos, solo el primer capítulo.

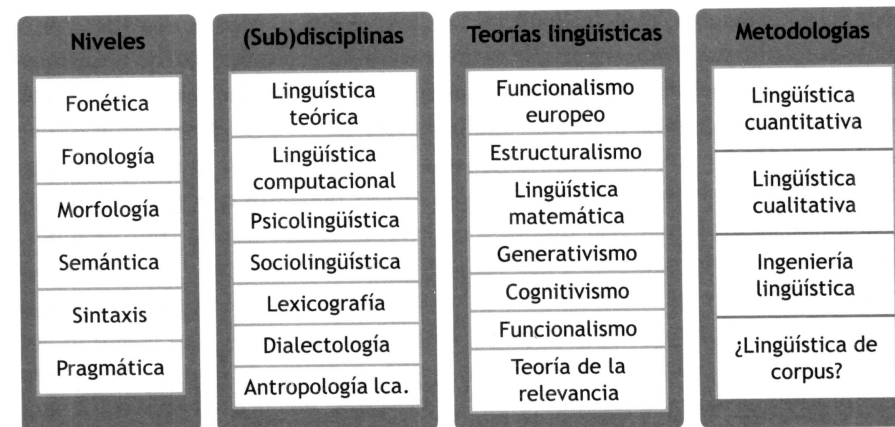
Además, se han tenido que compilar con una **intención concreta**, es decir, no son textos sin orden ni concierto, sino que buscan ser **representativos** de una

variedad lingüística en particular. Eso sí, esa variedad puede ser muy amplia: por ejemplo, un corpus que busca ser representativo del español escrito en el siglo XX o más concreta, textos que quieren ser representativos de las recetas de cocina escritas en español por anglófonos. El concepto de representatividad, aunque fácil de captar en su esencia, es difícil de determinar: ¿cuántos textos necesitamos para que algo sea representativo? Es por eso por lo que volveremos a ello en el capítulo 4.2.

2.2. LINGÜÍSTICA DE CORPUS

Ya hemos visto más de cerca qué es un corpus. Pero, entonces, ¿qué es la **lingüística de corpus**? La definición más fácil es la siguiente: la lingüística que hace uso de corpus. El problema surge cuando intentamos definir si es una metodología, una rama de estudio o una teoría lingüística.

Por la definición que acabamos de dar, la lingüística de corpus es una herramienta de trabajo, por lo que el mapa de la lingüística se parecería al que se puede observar en el esquema 2, en el que la lingüística de corpus aparece como una metodología. Se pueden usar corpus para hacer cualquier tipo de estudio lingüístico en cualquier nivel de lengua (semántica, sintaxis...), y, por lo tanto, la rama de estudio que le corresponde depende del fin para el que usemos el corpus. Pongamos que queremos hacer un diccionario, nuestra tarea pertenecería al campo de la lexicografía; si queremos extraer una teoría gramatical, estaríamos haciendo lingüística teórica; o pongamos que queremos entrenar el sistema de diálogo de un asistente virtual como Siri, estaríamos usando corpus para una tarea que pertenece a la lingüística computacional. En definitiva, trabajar con corpus es algo que podemos hacer en todas las subdisciplinas o ramas de la lingüística (se verá con más detenimiento en el capítulo 5). Esto hace que la lingüística de corpus sea más una metodología de trabajo que una rama de estudio.



Esquema 2. Encaje de la lingüística de corpus en las categorizaciones de la lingüística

Sin embargo, no todo el mundo está de acuerdo con esta categorización: los grandes pioneros de la lingüística de corpus y responsables de uno de los grandes hitos de los corpus que veremos más adelante, el corpus COBUILD (Firth, Sinclair y sus seguidores), defienden que la lingüística de corpus se ha ganado la categoría de marco teórico, ya que tradicionalmente las metodologías usadas en lingüística son más simples (McEnery & Wilson, 1996; Tognini-Bonelli, 2001).

Veamos por qué. En una investigación lingüística, se pueden usar métodos cuantitativos (dar porcentajes de aparición) o cualitativos (explicar el uso de una forma gramatical), usar estadística inferencial (establecer si la diferencia entre dos grupos de hablantes es estadísticamente significativa mediante pruebas estadísticas) o no, etc., pero todas estas metodologías o recursos no son únicas de la lingüística, sino compartidas con muchas otras ciencias sociales e incluso ciencias a secas. Los artículos de fonética, por ejemplo, contienen análisis estadísticos que se usan en biología para establecer parecidos entre genes y algunos análisis estadísticos de la sociolingüística tienen un diseño parecido al que se hace para probar si un fármaco funciona con un grupo objetivo al que le dan la medicación y otro grupo que recibe un placebo. El uso de porcentajes o de pruebas estadísticas no es algo único de la lingüística. Sin embargo, el uso de corpus es propio y único de la disciplina lingüística. La lingüística de corpus ha sido desarrollada por lingüistas y solo la usan ellos.

Además, las metodologías tienen que ver con el tipo de análisis que se hace de los datos y su posterior tratamiento estadístico, mientras que la lingüística de corpus

tiene relación con los datos. La lingüística de corpus es los datos. A esos datos, una vez conseguidos mediante corpus, se les puede aplicar cualquier tipo de análisis, es decir, cualquier metodología. Lo podemos ver con un ejemplo. Pongamos el caso de que podemos conseguir las producciones escritas (redacciones) de un grupo de alumnos de A1 y A2 que han hecho una descripción de sí mismos. Una vez conseguidas, tendremos un corpus. Después de eso, podemos aplicar 1) un análisis cualitativo, es decir, explicar qué tipo de errores han cometido (por ejemplo, errores en el uso del presente o que han confundido el verbo *ser* y el *estar*); 2) un análisis cuantitativo, por ejemplo, decir cuántos errores de *ser/estar* han cometido o qué tanto por ciento de las veces que han usado *ser/estar* lo han hecho de un modo erróneo; y 3) un análisis estadístico, hacer un conteo de los errores de cada tipo (morfológico, sintáctico, de número...) y aplicar una prueba estadística para comprobar si los alumnos de A1 han hecho significativamente más errores que los alumnos de A2. Todas estas metodologías son compatibles e independientes de que usemos o no corpus en nuestro estudio.

Otra de las particularidades de la lingüística de corpus es que muchas teorías lingüísticas, como el cognitivismo y la gramática de construcciones, han surgido gracias a ella, y se han desarrollado a su sombra. Incluso ramas de la lingüística enteras como la lingüística computacional deben su existencia y comparten toda su historia y evolución con la lingüística de corpus. Todo ello hace que muchos autores prefieran entender la lingüística de corpus como una rama más de la lingüística.

Pero, para el objetivo que aquí nos ocupa, lo más adecuado es entender la lingüística de corpus como una **herramienta** que va a ser de ayuda para conocer mejor el perfil de nuestros alumnos, crear mejores materiales para ellos, poner ejemplos más realistas, y proponer dinámicas de clase divertidas. Por lo tanto, lo más cercano de todas las opciones propuestas, es definir la lingüística de corpus como una metodología de trabajo, pero siempre teniendo en cuenta que no es una metodología al uso.

2.3. LOS PRIMEROS CORPUS

Ya hemos hablado de los primeros lingüistas que recogieron datos, aunque, esos son solo precursores. Tanto ellos como los antropólogos contaban con muestras de lengua, pero en forma de atlas, no de corpus, ya que sus muestras no estaban precisamente ordenadas ni eran digitalizables. Entonces, ¿cuándo se crean los primeros corpus?

Como hemos dicho, la historia de la lingüística de corpus está ligada a la posibilidad de digitalizar los datos, y esto es decir que está ligada a la historia de los ordenadores. Y, como ella, ha sufrido grandes cambios desde los años cincuenta del siglo XX cuando apareció el primer ordenador.

Algunos años antes, los estructuralistas habían abandonado la rama normativista de la lingüística y querían describir la lengua que se hablaba en la calle. Uno de sus objetivos era crear diccionarios y gramáticas que se basaran en un uso real de la lengua con ejemplos documentados en textos. A esa tendencia pertenece Hornby (Hornby et al., 1948), que ideó el diccionario en el que se inspiró María Moliner para crear su famoso *Diccionario de Uso de la Lengua Española* (Moliner, 1966), todavía compilado a mano, con fichas de biblioteca. La aparición de esas primeras obras descriptivistas llevó a un gran conjunto de lingüistas a percibir la necesidad de contar con grandes bases de datos de textos que estuvieran almacenadas y ordenadas, y qué mejor para esa tarea que informatizar las bases de datos de lengua.

A la vez que pasaba todo lo anterior, nació el primer centro del mundo dedicado a la lingüística de corpus. Se trata del *Survey of English Usage* del University College London (SEU) que todavía sigue activo en la actualidad y, como su nombre indica, trababa de documentar el uso real del inglés, y lo hacía a partir de textos. Sus miembros compilaron el primer corpus de un millón de palabras, en aquella época todo un hito, bajo la coordinación de Randolph Quirk. Los resultados obtenidos de ese corpus llevaron a la redacción de *A Comprehensive Grammar of the English Language* (Quirk et al., 1985). Otros clásicos de la época que sirven para comprender el inicio de los corpus son el corpus Brown, el Lancaster-Oslo/Bergen Corpus (conocido como LOB) o el *Kolhapur Corpus* (con datos de inglés americano, británico, y de la India).

A partir de los años ochenta, las nuevas capacidades de los ordenadores permitieron compilar corpus más abundantes. Entre ellos, probablemente el más famoso sea el *Bank of English* (corpus usado para crear el diccionario COBUILD) de 450 millones de palabras e impulsado originalmente por John Sinclair, toda una referencia en los estudios de corpus (Sinclair, 1987). El COBUILD, como el diccionario de Hornby, estaba pensado especialmente para estudiantes de lengua extranjera; los estudios basados en corpus se aplicaron al estudio de segundas lenguas desde muy pronto.

Otros corpus del inglés que hicieron aportaciones importantes al campo son el *British National Corpus*, de 100 millones de palabras y que contiene muestras de

habla oral y escrita, y el *American National Corpus* (ANC), un corpus oral de inglés norteamericano (Ide & Macleod, 2001). Sin embargo, en este libro no nos vamos a centrar en las aportaciones de cada uno de ellos, basta con saber que existieron, y que cada uno fue aportando, a través de sus aciertos y errores, su pequeña contribución para crear la disciplina que conocemos en la actualidad.

2.4. LOS CORPUS DEL ESPAÑOL

En el caso del español, la historia es algo más reciente. Existieron primero algunas iniciativas de digitalización de textos antiguos, como la realizada para la redacción del *Dictionary of Old Spanish Language* (Nitti, 1979) y también encuestas como las del *Proyecto de estudio coordinado de la norma lingüística culta del español hablado en las principales ciudades de Iberoamérica y de la Península Ibérica*, un proyecto que fue propuesto inicialmente en 1964. Sin embargo, este proyecto, comparable al *Survey of English Usage*, no se podría considerar propiamente un corpus hasta algo más tarde, cuando se digitalizó y clasificó porque, como hemos dicho, los corpus propiamente dichos nacen con los ordenadores y en España eso es entre los años 1980 y 1995.

Pero dejemos de lado la parte más histórica y pasemos a los corpus del español que se crearon desde los años 90 del siglo pasado y que siguen disponibles hoy en día para su consulta. En esta sección, hablamos solo de corpus de habla nativa, que pueden servir para dar ejemplos positivos y para que los estudiantes puedan hacer búsquedas.

A los corpus de español hablado como L2 le dedicaremos un capítulo aparte (el 6) porque tanto su método de consulta como su finalidad son diferentes. Los corpus que contienen español hablado por no nativos son corpus pensados para el investigador que quiere saber más de la interlengua de sus estudiantes (Selinker, 1972), que quiere ver qué errores cometen. Por lo tanto, son corpus para la investigación del ELE. Mientras que los corpus de lengua nativa, en su mayoría, no están pensados para una tarea específica; pertenecen a aquella clase que llamaremos **corpus generales**.

Pese a esto, algunos de los corpus de habla nativa más grandes sí han sido diseñados con un objetivo en mente. Los corpus más grandes de los años noventa fueron fruto de la inversión de empresas editoriales que, como veíamos en el caso del COBUILD, fueron usados para crear diccionarios de nueva planta. Así se financiaron, por ejemplo, el *Corpus Vox-Bibliograf* (CVB) (10 millones de formas) o el CUMBRE de SGEL (20 millones) o el *Corpus del español mexicano contemporáneo* usado para el *Diccionario del español de México* (CEMC, 2000 formas). En el campo de los corpus orales, destaca el *Corpus Oral de Referencia de la Lengua Española*

Contemporánea (CORLEC) una iniciativa de los años noventa que cuenta con un millón de formas orales. También se financiaron corpus para uso médico: en el campo de la lingüística clínica fue especialmente importante el corpus LEXESP, usado para elaborar el *Diccionario de frecuencias de las unidades lingüísticas del castellano* a partir de 1,5 millones de formas (Alameda & Cuertos, 1995).

Estos corpus, aunque muchos están disponibles, no se han pensado con el objetivo de facilitar la búsqueda en ellos y, por lo tanto, no cuentan con una interfaz en línea en la que podamos hacer búsquedas de manera cómoda. Es decir, su disponibilidad consiste en que puedes descargar todos sus textos en tu ordenador, pero cómo busques en ellos, es cosa tuya. Las búsquedas deben realizarse mediante un buscador que creen los investigadores. Aprender a hacer esto requiere un tipo de formación que no tiene ninguna relación con el ELE. Quizá, si hiciste lingüística de corpus o computacional en la carrera, sepas hacerlo; en ese caso, puedes echarles un ojo. Sin embargo, en este libro nos centraremos en los corpus que tienen una interfaz para facilitar la búsqueda al público general.

Entonces, ¿cuáles son los corpus que podemos consultar mediante una interfaz? Ahora los vemos. Los primeros corpus de habla nativa en español que podemos encontrar en línea son compilaciones de textos u obras que realizan diversas universidades para fines de investigación. De ellos, el más usado es probablemente el Spanish On Line o corpus SOL (Mighetto, 2003), consultable en línea desde 1998 (figura 1). En él se pueden hacer búsquedas por palabras o secuencias para obtener en qué contextos se usa la palabra o cuál es su frecuencia de uso.

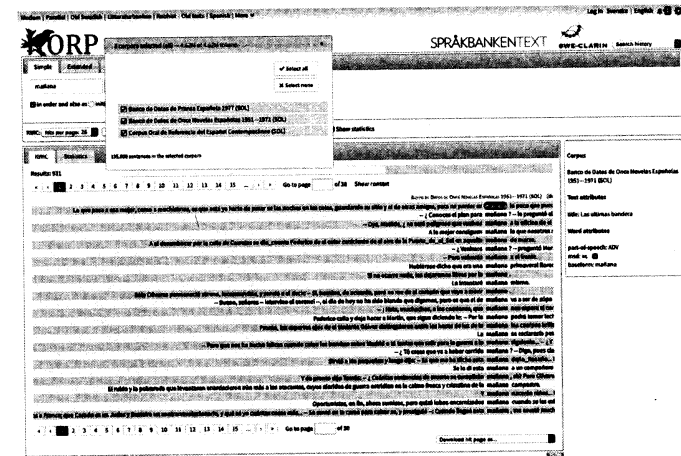


Figura 1. Interfaz del corpus SOL (Mighetto, 2003)

Pero, los mejores corpus para nuestro uso son aquellos que nacen de grandes iniciativas. Y los primeros de ellos son los corpus de referencia compilados por las academias de la lengua. Los corpus de referencia tienen el objetivo, como en su día el *Vox-Biblograf*, de servir como base de datos para actualizar un diccionario (por eso de referencia), en este caso el *Diccionario de la Lengua Española* (DRAE) y el *Nuevo diccionario histórico del español* (NDHE). Pero sus creadores también los han dotado de una interfaz de búsqueda en línea que permite a los usuarios hacer búsquedas sencillas para consultar la frecuencia de una palabra. Los primeros de estos corpus fueron el *Corpus de referencia del español actual* (CREA), que data de 1995, y el *Corpus diacrónico del español* (CORDE), cuya compilación se aprobó poco después del CREA. Probablemente, estos fueron los corpus sincrónicos (CREA) y diacrónicos (CORDE) más usados hasta la irrupción del llamado *Corpus del español* (CdE) (Davies, 2002).

El CdE marcó un antes y un después en el uso de los corpus, porque usa una interfaz de búsqueda (la del BYU) que había tenido mucho éxito en inglés. Esto significa que permite búsquedas de secuencias de palabras mucho más refinadas que el CREA y el CORDE. Por ejemplo, en él se pueden buscar frases que contengan la conjunción *si* seguida de un verbo en subjuntivo, o los sustantivos que acompañan más frecuentemente a la palabra *tomar*. Lo veremos en los ejercicios del capítulo. Además, el CdE cuenta con varios subcorpus: el *Genre/Historical*, que contiene textos de novelas hasta el 2000; el *Web/Dialects*, que podemos considerar el corpus más general; el NOW, un corpus que contiene noticias y textos periodísticos publicados en internet (este último es lo que llamamos un corpus monitor, lo veremos en el capítulo 4); y *n-grams* (BYU), que contiene libros obtenidos de Google Books. Por todo ello, el éxito del CdE fue inmediato y todavía hoy es el corpus más versátil y el que más usaremos en los ejercicios. La única pega es que esa versatilidad (la posibilidad de refinar mucho nuestras búsquedas) implica que buscar en él es un poco más complejo que en el resto de los corpus.

Volviendo a los corpus académicos, el CREA se quedó corto enseguida, tanto por su contenido (mayormente textos literarios y periodísticos casi al modo del *Diccionario de Autoridades*¹) como por su interfaz. Por eso, hoy contamos con un nuevo corpus heredero del CREA, el CORPES, que permite búsquedas casi igual de refinadas que el *Corpus del Español*. Este será el segundo corpus que más usaremos.

1 El *Diccionario de Autoridades* es el primer diccionario publicado por la Real Academia Española. En él solo se incluían palabras que habían sido documentadas en autores célebres de la literatura española.

2.5. INTERNET COMO CORPUS Y GOOGLE BOOKS

Mientras explicábamos los primeros corpus y los corpus de los años noventa, hablábamos de corpus de millones de palabras. La primera y gran barrera que se superó en términos de número de palabras fue mil millones de palabras (un *billion* en inglés). A primera vista, esto puede parecer mucho, pero en realidad es una muestra pequeñísima comparada con la cantidad de palabras que obtendríamos si sumamos, por ejemplo, todas las publicaciones de los periódicos durante un mes. Solo para dar una idea, en la actualidad, un corpus como el COBUILD para el inglés contiene 4,5 mil millones de palabras. El CORPES contiene, en 2019, 225 millones de palabras y tiene como objetivo alcanzar los 400 millones. Mientras que el CdE contiene, para sus cuatro subcorpus: 100 millones de formas en el histórico, 2000 millones de formas en el web, 5500 millones de formas en el NOW, y 45 mil millones en *n-grams* (BYU).

Pero en los tiempos que corren, nos estamos dejando un corpus primordial. ¿A quién preguntan los estudiantes cuando tienen dudas con un régimen verbal? Por supuesto, a Google. Google no es un corpus en sí mismo, es un buscador, como la interfaz de búsqueda en los corpus. Estrictamente, internet no es un corpus, es un conjunto de datos digitalizado, sí, pero no ordenado, más bien lo contrario, desordenado y caótico, ahí reside su gracia. En internet no sabemos (siempre) quién es el autor del texto o los años que tenía cuando lo escribió, o ni siquiera si es hablante de lengua nativa o extranjera. Pero, pese a todos estos problemas, internet nos da acceso a un gran banco de datos que se han aprovechado de innumerables maneras.

Existe bibliografía sobre cómo usar internet como corpus y también trabajos que han usado como corpus la Wikipedia, Twitter o Amazon. Sin embargo, hacer búsquedas refinadas en esos servicios no es fácil. Los investigadores que lo hacen tienen que crear sus propios buscadores que diseñan en lenguajes de programación como Python o R. o bajar los textos e introducirlos en un software gestor de corpus que ayuda a ordenar los textos, analizarlos y realizar búsquedas, como el software libre AntConc (Anthony, 2019) o SketchEngine, de pago, pero con una versión de prueba gratuita (Kilgariff et al., 2014). En este libro, no explicaremos cómo usar estos programas o cómo realizar análisis propios en, por ejemplo, Python. Si fueras un lingüista computacional, o estuvieras cursando un máster de Humanidades Digitales, lo necesitarías, pero nosotros vamos a usar corpus a nivel usuario.

En cambio, existe una porción de internet que cuenta con su propio buscador y que puede ser muy útil en trabajos de cultura y ELE: se trata de Google NGrams

Viewer, un motor de búsqueda que trabaja con un conjunto de libros extraído de Google Books (45 000 millones de palabras). Para tener una idea de lo que se puede hacer con un corpus de tantos libros, recomendamos ver una *TED talk* muy amena que presenta Culturomics, un proyecto para explicar el cambio cultural a través de los textos. La charla se llama “¿Qué hemos aprendido de 5 millones de libros?” y está disponible de manera gratuita en la página de TED Talks y en su canal de YouTube.

Los corpus que hemos mencionado en este capítulo no son ni mucho menos todos los disponibles para el español; para una lista más exhaustiva puedes consultar Cruz Piñol (2012), pero nosotros, en este manual y en sus actividades, vamos a usar los que consideramos más útiles y fáciles de manejar para ELE, principalmente: el CORPES, el *Corpus del español*, la interfaz de búsqueda de Google Books y Google NGram Viewer.

2.6. PARA MÁS INFORMACIÓN

Rojo, G. (2016). Los corpus textuales. En: Gutiérrez-Rexach, J. (ed.), *Enciclopedia lingüística hispánica*. Oxon: Routledge, 285-296.

Michel, JB. & Lieberman Aiden, E. (Julio, 2011). What we learned from 5 million books. Conferencias TED. https://www.ted.com/talks/jean_baptiste_michel_erez_lieberman_aiden_what_we_learned_from_5_million_books.

2.7. ACTIVIDADES

2.7.1. Fuera del mundo de la lingüística la gente no suele pensar en datos o muestras de habla cuando decimos *corpus*. Vamos a comprobar esta intuición en un corpus de habla real, el *Corpus del Español* (CdE) (Davies, 2001). Para ello, vamos a usar un corpus online con interfaz de búsqueda. Es decir, un buscador parecido a Google pero con textos ordenados. ¡Atención, el corpus de verdad son las muestras de habla que hay detrás, el buscador es solo una manera de consultar los datos de forma más fácil! El CdE es gratuito, pero después de un número de búsquedas pide que te registres, ¡no dudéis en hacerlo! El registro para uso de profesores o estudiantes es gratuito y lo vamos a usar mucho. Antes del CORPES, este era el único corpus con buscador lematizado para el español (veremos qué significa esto en el siguiente capítulo), eso hace que haya una infinidad de trabajos realizados con él.

1. Ve a la dirección web: <https://www.corpusdelespanol.org/web-dial/>
2. Escribe en el cuadro de búsqueda *corpus* y pulsa el botón FIND MATCHING STRINGS. A continuación, aparece la frecuencia de la palabra “FREQ”. ¿Cuántas veces aparece la palabra *corpus* en el corpus web del CdE?
3. Pulsa en la palabra *corpus*. Se abrirá una nueva ventana con el contexto en el que aparece la palabra. Verás que hay 92 páginas de resultados, eso es porque había muchas ocurrencias de esa palabra. A partir de la primera página de resultados, ¿podrías decir cuál es el significado más habitual de la palabra *corpus* en español?
4. Ahora compara los resultados con un corpus de inglés, el COCA (<https://www.english-corpora.org/coca/>), ¿cuál es el significado más habitual en inglés americano?

2.7.2. Ve al corpus SOL (<https://spraakbanken.gu.se/korp/?mode=spanish#?lang=en&cqp=%5B%5D&page=0&search=word%7Cemail>) y busca las siguientes palabras en sus tres subcorpus: *presidente*, *presidenta*, *internet*, *email*.

1. ¿Cuáles aparecen? ¿Por qué crees que no aparecen algunas?
2. En la información de los subcorpus, busca los años de los textos que contiene cada uno. ¿Cuál de los dos es más útil para un alumno de ELE?

3

CÓMO HABLAR DE CORPUS: CONCEPTOS CLAVE Y TERMINOLOGÍA EN LINGÜÍSTICA DE CORPUS

3. CÓMO HABLAR DE CORPUS: CONCEPTOS CLAVE Y TERMINOLOGÍA EN LINGÜÍSTICA DE CORPUS

Antes de continuar con la teoría de lingüística de corpus, vamos a introducir algunos conceptos clave de la disciplina que servirán para hablar con propiedad (con la terminología adecuada) durante el resto del libro. En los ejercicios del capítulo anterior, han aparecido algunos de ellos, pero ahora se van a sistematizar. Para seguir mejor las explicaciones es recomendable haber hecho (o, para el simple lector, haber leído) el ejercicio 2.5.1.



Esquema 3. Mapa del contenido del tema

3.1. FORMAS Y LEMAS

Los corpus contienen palabras, pero el problema es que la palabra *palabra* es bastante escurridiza y difícil de definir. Por ejemplo, es discutible si el *lo* de *piénsalo* es una palabra diferenciada de *piensa* o si juntos forman una nueva palabra. En general, en lingüística se pueden tomar diferentes criterios para delimitar las palabras: fonéticos, morfológicos o funcionales, principalmente (González Calvo, 2000). Para no tener que entrar en ese debate, en corpus se definen como aquello

que queda entre dos espacios, es decir, con un criterio meramente gráfico. A esas “palabras entre dos espacios” se las llama **formas**. Además, sabemos que las palabras flexionan: por ejemplo, *piensa* y *pienso* son formas flexionadas de una misma palabra; también *amiga* y *amigo*, y por eso se agrupan bajo la misma entrada en el diccionario. Si queremos encontrar en el diccionario el significado de *amiga*, buscaremos en la entrada de *amigo*. Eso en corpus es un **lema**. El lema *amigo* contiene las formas *amigo/a/os/as*, *amiguísimo/a/os/as*, *amigote/s*, etc.

3.2. OCURRENCIAS

Cada una de las apariciones de una palabra o, mejor dicho, de un lema en el corpus es una **ocurrencia** (*token*). Si una palabra aparece muchas veces en el corpus, decimos que hay muchas ocurrencias de ese lema o de esa forma del lema. Por norma general, cuando hacemos una búsqueda en un corpus, el sistema devuelve cada una de sus ocurrencias; es decir, cada vez que la palabra (sea lema o forma) se ha documentado en el corpus.

3.3. FRECUENCIA

La **frecuencia** de una palabra en un corpus es su número de ocurrencias, es decir, el número de veces que esa palabra aparece en el mismo. La frecuencia puede ser absoluta o normalizada.

La frecuencia absoluta es un valor muy útil para saber si una palabra se usa mucho o poco, pero no es lo mismo que la palabra *estudiante* aparezca cinco veces en una redacción de 500 palabras a que aparezca cinco veces en la novela de un autor que sea de 5 000 000. Su frecuencia es la misma, cinco, pero la “densidad”, no. En la redacción ha aparecido mucho más y probablemente se trate de uno de los temas del texto, mientras que *estudiante* aparezca cinco veces en una novela no nos dice mucho de los temas de la obra (más que quizá en la novela haya un hijo o un amigo estudiantes).

Para que la frecuencia nos diga algo más sobre si la palabra aparece mucho o poco, si tenemos en cuenta el tamaño del corpus en el que aparece, existe la **frecuencia normalizada**. En los dos casos expuestos arriba, la frecuencia es la misma (5), pero no la frecuencia normalizada. En un caso, *estudiante* representa el 1 % del corpus (1 palabra de cada 100 de las que aparecen), mientras que en el segundo caso constituye el 0,0001 % del corpus. La frecuencia normalizada se calcula dividiendo el número de ocurrencias entre las palabras totales del corpus y, normalmente, se expresa en porcentaje, como hemos hecho aquí. La

fórmula matemática para expresarlo sería la siguiente (no es más que la notación matemática de un porcentaje):

$$\text{Frecuencia normalizada en porcentaje} = \left(\frac{N \text{ de ocurrencias}}{\text{tamaño del corpus}} \right) \cdot 100$$

Fórmula 1. Frecuencia normalizada en porcentaje

Pero existe otra manera (muy parecida, por otro lado) de expresar la frecuencia de una palabra en un corpus. En vez de hacerse por cada 100 palabras (porcentaje), se calcula por cada un millón de palabras. ¿Y por qué un millón? Por una mera casualidad histórica. Durante una época, muchos corpus eran de un millón de palabras y así, se podían ahorrar el paso de multiplicar, puesto que, en un corpus de un millón de palabras, para calcular la “frecuencia por millón” habría que dividir entre un millón y después multiplicar por un millón, o lo que es lo mismo, no hacer nada. ¿Entonces por qué se ha mantenido esta normalización vigente hasta ahora? Por una cuestión práctica. Si se calcula el porcentaje en un corpus grande, los números resultantes son muy pequeños y quedan cosas como 0.0000001 %, un tipo de números que son poco prácticos para escribirlos o recordarlos. Si seguimos con el ejemplo de arriba, nuestras cinco ocurrencias de *estudiante* en un corpus de 5 000 000 de palabras, la frecuencia normalizada en porcentaje es de 0,0001 %, pero por millón de palabras sería 1 %: la palabra *estudiante* aparece una vez por cada millón de palabras. Esta cifra es mucho más intuitiva que la frecuencia en porcentaje.

$$\text{Frecuencia normalizada por millón de pal.} = \left(\frac{N \text{ de ocurrencias}}{\text{tamaño del corpus}} \right) \cdot 1\,000\,000$$

Fórmula 2. Frecuencia normalizada por millón de palabras

El tipo de frecuencia normalizada que usa un corpus depende de las decisiones que se tomaron en el diseño de este y siempre se indicarán en la documentación o en la página web del corpus. Por ejemplo, en el CORPES, al situar el ratón encima de *Fnorm* cuando se busca una palabra, un cuadro explícita “N de ocurrencias por cada millón de palabras en cada zona”, pero en Google Ngrams (otro corpus que se usará en este libro) la frecuencia se expresa en porcentaje.

3.4. CONCORDANCIAS

Después de ver la frecuencia de la palabra, el sistema de búsquedas ha devuelto cada una de las ocurrencias de la palabra en el corpus. Pero no la ha devuelto sola, nos la ha devuelto en contexto (y esta es la gracia de los corpus). A esto lo llamamos líneas de concordancias, o simplemente **concordancias** (en inglés, *Keyword In Context – KWIC*). En la figura 2, hay un ejemplo de KWIC. Si buscamos el lema *aprendiente* en el corpus CORPES, el sistema devuelve: 1) cada ocurrencia (formas *aprendiente* y *aprendientes*) que hemos buscado destacada y centrada, en este caso, se destaca en azul y negrita, pero que en otros corpus destacan la forma buscada de otras maneras; y 2) el contexto por delante y por detrás de donde aparece.

que se preparen cursos llamados de orientación, con el fin de sensibilizar a los aprendientes	a las condiciones específicas en las que se ofrece la educación a distancia. Generalmente
las que se ofrece la educación a distancia. Generalmente dan énfasis al papel del aprendiente	y del tutor/asesor, y pueden tratar también formas de organizar el tiempo y hábitos
Lingüística Aplicada a Distancia? es familiarizar a los profesores y los aprendientes	en el campo de la lingüística aplicada, por ser éste un campo en el que se han
módulo pretende además problematizar el contenido y despertar la curiosidad de los aprendientes	En este sentido se intentó seguir el esquema, que para el diseño de una guía didáctica
caso o producto integrador, pues de esta manera las actividades preparan a los aprendientes	para realizar ese producto final. Después, es importante diseñar una actividad
producto final. Después, es importante diseñar una actividad preliminar que ubique al aprendiente	en el nivel inicial en que se encuentra (*) En la versión preliminar del módulo
esquemas de comunicación entre los participantes y las tutoras, y entre los mismos profesores/aprendientes	mediante videoconferencia, un foro de discusión electrónico, el correo electrónico
tutoras dieran información general o hicieran aclaraciones pertinentes, pues todos los profesores/aprendientes	reciben, al mismo tiempo, los mensajes. Estos mensajes fueron numerados progresivamente
ellas merecían aspectos importantes en sus análisis y comentarios. En general, los aprendientes/mestros	expresaron respeto y admiración por las tutoras, debido a que además de ser pacientes
análisis conceptual. En los comentarios que surgieron a esta pregunta, algunos de los aprendientes/mestros	sugirieron que se sirva a los participantes un mejor manejo de la tecnología, para
En el aspecto del desempeño individual, la mayoría de los maestros/aprendientes	(71%) estuvieron de acuerdo con la sugerencia de una elaboración de horario. Afirmaron

Figura 2. Concordancias del lema *aprendiente* en el corpus CORPES

3.5. COLOCACIONES

Relacionado con el concepto de concordancias, existe el concepto de *colocación*, que quizá conozcas de diferentes estudios de gramática, si has estudiado una filología, o de los enfoques léxicos en ELE. De hecho, de los conceptos vistos hasta ahora el de colocación es el único que ha sido explotado en ELE gracias al enfoque léxico.

Una **colocación** es una secuencia de palabras que aparecen juntas. En términos teóricos, podría ser cualquier tipo de secuencia desde frases lexicalizadas (y el estadio más avanzado sería un refrán) hasta secuencias de palabras que aparecen en un orden concreto. Por ejemplo, en español decimos *blanco y negro*, mientras que en inglés se dice *black and white*, que se traduciría literalmente por “negro y blanco”. Pero en lingüística de corpus se contemplan principalmente colocaciones formadas por dos o tres palabras.

En español, son especialmente productivas aquellas formadas por:

- un nombre + preposición + nombre: *punto de encuentro, punto de vista, plan de pensiones...*
- un nombre + adjetivo: *horas intempestivas, nariz aguileña...*

3. CÓMO HABLAR DE CORPUS: CONCEPTOS CLAVE Y TERMINOLOGÍA EN LINGÜÍSTICA DE CORPUS

- verbo + nombre: *ponerse las pilas, plantear problemas...*
- verbo + preposición + sustantivo: *llevar a cabo, ponerse de acuerdo...*

Aunque estas no son las únicas. A su vez, un **colocado** o **colocativo** es la palabra que aparece habitualmente en una colocación como complemento. Por ejemplo, en la colocación *plantear problemas*, *problemas* es un colocado de *plantear*, porque a) complementa a *plantear* y b) juntos forman una colocación. Así, en algunos corpus, para buscar colocaciones se busca una palabra y después se buscan sus colocados a una distancia de cero palabras (como en el caso de *plantear problemas* en que *problemas* aparece inmediatamente después de *plantear*), de una palabra (*punto de encuentro*, *encuentro* aparece a una palabra de distancia de *punto*) e incluso hasta una distancia de seis palabras. Entonces, ¿cómo decidimos qué es una colocación y qué no?

Para decidir si algo merece ser considerado como colocación, medimos su fuerza calculando el número de veces que las palabras aparecen juntas respecto de las que aparecen por separado. Esa medición de la “fuerza de la colocación” se realiza mediante el llamado **criterio de costo** (*cost criteria*) que consiste en una fórmula, que no veremos aquí, creada por Kita et ál. (1994). Esta fórmula se creó precisamente para determinar qué colocaciones de un corpus sería útil enseñar en clase de lengua extranjera. Es gracias a ella y a sus adaptaciones que los corpus nos pueden ofrecer una lista finita de las colocaciones que aparecen en él. Si no existiera un filtro para decidir qué es una colocación, cualquier palabra podría ser colocado de otra en el momento en el que aparecieran juntas una vez, como *aparecer juntas*, y eso sería muy poco informativo.

Por lo tanto, dependiendo de su criterio de costo, las colocaciones pueden ser más fuertes o menos. Es decir, hay palabras que se combinan muy frecuentemente con otras y solo con esas. Casos de estas serían: *módico precio, horas intempestivas, pingües beneficios, error garrafal*; donde el adjetivo *módico* solo se usa para modificar a *precio*; en español, lo único que es intempestivo son las horas; lo único que es pingüe son los beneficios; y lo único que es garrafal son los errores. En cambio, otras tienen una fuerza más limitada. *Blanco y negro* es una colocación y como tal estas palabras aparecen con mucha frecuencia juntas, pero también usamos con mucha frecuencia las dos palabras por separado, ya que podemos decir que un coche es blanco o que es negro sin problemas.

Hoy en día existen al menos dos manuales de ELE que hacen un uso extenso de las colocaciones en clase. El primero, de la editorial Difusión, se presentó en 2011: se trata del manual “Bitácora”, disponible de A1 a B2, y las integra en la clase

como "Palabras en compañía". El segundo, *Frecuencias* de Edinumen, se presentó en 2020 y se ha creado a partir del corpus CORPES. Otros manuales que incluyen colocaciones, pero no están basados únicamente en ellas, son *Aula internacional*, *Gente hoy*, *Gente joven*, *C de C1*, *Campus Sur* y *Reporteros internacionales* (Difusión) y *Nuevo Prisma* de Edinumen en el nivel C.

También gramáticas como la *Gramática comunicativa del español* (Matte Bon, 1995) usan de manera indirecta las colocaciones cuando pueden sacar provecho de ellas para explicar de manera más sistemática la gramática. Esto no nos debería extrañar si recordamos que el primer propósito de la creación de corpus fue precisamente la creación de diccionarios y gramáticas basados en el uso. Si tienes curiosidad sobre las colocaciones, puedes consultar en línea el *Diccionario de Colocaciones del Español* (DICE).

3.6. CORPUS ANOTADO

La ventaja de los corpus es que son conjuntos de datos ordenados y estructurados: cuanta mayor información se incluya de las palabras, podremos hacer búsquedas más refinadas. Por ejemplo, en un corpus que contenga errores de estudiantes de nivel B1, puede ser útil tener los datos ordenados por L1 de los aprendientes o por tipo de error. Así, podríamos obtener los errores de orden de palabras más habituales en sinohablantes de nivel A2. Esto es lo que llamamos un **corpus anotado**. Los niveles de anotación pueden ser múltiples, podemos anotar desde la categoría morfológica de las palabras, o los papeles temáticos (a veces llamados **roles semánticos**, como en el corpus ADESSE que veremos en las actividades de este capítulo) hasta, como decíamos, el tipo de error o si una grabación contiene risas o si un manuscrito original estaba rasgado. Se verá más en detalle en el capítulo siguiente.

3.7. CORPUS LEMATIZADO

Los corpus pueden estar **lematizados**. Es decir, que para cada **forma** de cada palabra (**variante**) alguien ha anotado a qué lema (a qué entrada del diccionario) pertenece esa forma. De esta manera si buscamos en un corpus el lema *vivir*, aparecerán todas las concordancias de formas de ese lema como *Yo vivo en Canarias* o *Yo vivo en Venezuela*, pero también *¿Dónde vives?*.

Si un corpus no está lematizado, o si prefiero buscar por forma, y realizo la consulta *vivir*, solo aparecerán aquellos contextos en que *vivir* se ha usado en infinitivo como *Se fue a vivir a los EE.UU.*

3.8. PARA MÁS INFORMACIÓN

Villayandre Llamazares, M. (2008). Lingüística con corpus (I). *Estudios Humanísticos. Filología*, (30), 329-349.

3.9. ACTIVIDADES

3.9.1. Ve a la web del CORPES (<http://web.frl.es/CORPES/view/inicioExterno.view>).

1. Consulta la estadística de una palabra gramatical, por ejemplo, *de* y de una palabra léxica, por ejemplo, *amigo*. Anota la frecuencia relativa de cada una de las palabras que hayas escogido. ¿Cómo está normalizada la frecuencia? ¿Por millón? ¿En porcentaje? ¿Por qué crees que la palabra gramatical es más frecuente?
2. Busca forma y lema de *amigas* en el CORPES y observa los resultados. ¿Por qué *amigas* no se puede buscar como lema? ¿Cuál sería entonces el lema de *amigas*?
3. Si quiero obtener información sobre el uso de la palabra *jueza* en los últimos años, ¿tendría que buscar por forma o por lema?

3.9.2. Ve al Corpus Diacrónico del Español (CORDE) y realiza una consulta de la palabra *campera*.

1. Pincha en VER ESTADÍSTICA para ver la frecuencia, ¿está normalizada?
2. Vuelve atrás y pincha en OBTENCIÓN DE EJEMPLOS > RECUPERAR. Fíjate en los textos donde se ha encontrado. ¿De dónde es típica?
3. ¿Crees que el corpus CORDE está lematizado?

3.9.3. Busca en el CORPES las colocaciones (en el CORPES las llaman *coapariciones*) de categoría verbo de la palabra *posibilidad*.

1. ¿Cuáles son las colocaciones V+ N (verbo seguido de nombre) más frecuentes? Para hacerlo, en el menú superior, selecciona COAPARICIONES. En clase de palabra, selecciona NOMBRE. Para ordenar los resultados, pincha en los encabezados de la tabla de resultados, por ejemplo, FREQ, CLASE.

3.9.4. Ahora vamos a buscar colocaciones en el Corpus del Español (CdE) en su subcorpus *web-dialects*.

1. ¿Cuáles son los colocados nominales más habituales del verbo *querer*? Si buscas *querer* en minúsculas, solo aparecerán formas en infinitivo. Si quieres hacer una búsqueda por lema, escribe *QUERER* en mayúsculas. Para encontrar los colocados, selecciona en la parte superior del cuadro de

búsqueda COLLOCATES y en la banda con números, selecciona 0 COLLOCATES por la izquierda de la palabra y 2 COLLOCATES por la parte de la derecha. ¿Por qué 2 y no 1? Porque el sistema cuenta palabras, así que en una estructura como “querer la luna” el nombre, luna, está a dos posiciones de distancia. Elige NOUN en la categoría gramatical (PART OF SPEECH, POS) del colocado y verbo en la del lema que se buscará.

2. ¿Para qué puede servir en una clase de ELE saber los colocados más frecuentes de *querer*?

3.9.5. Ve al formulario de búsqueda avanzada de ADESSE (<http://adesse.uvigo.es/data/avanzado.php>). Busca por rol semántico y selecciona el siguiente rol: SENSACIÓN.A1-EXPERIMENTADOR. Pulsa en REALIZAR BÚSQUEDA y, después, VERBOS. ¿Por qué podría ser interesante presentar esa lista de verbos juntos en clase?

4

TIPOLOGÍA Y DISEÑO DE CORPUS

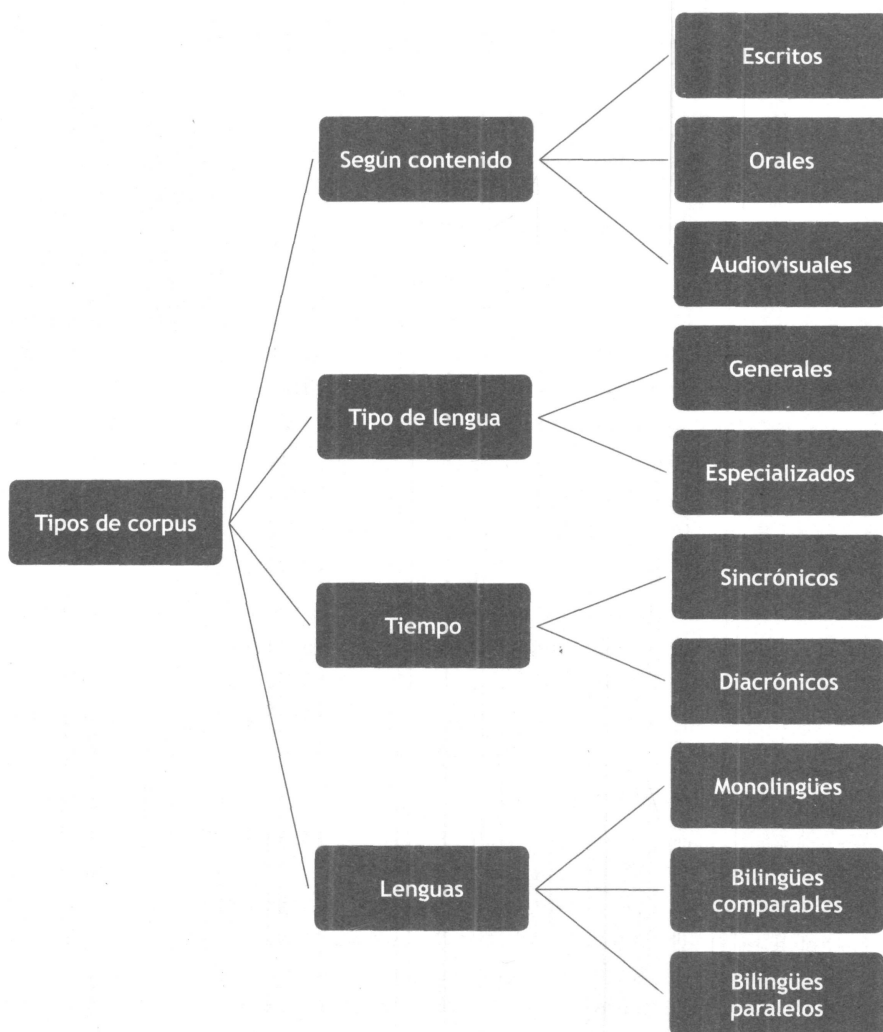
Este capítulo trata de la creación de corpus. En él se habla de las decisiones técnicas que hay que tomar antes de empezar a crearlo y de los pasos que se siguen hasta tenerlo completo. Durante todo el capítulo, se hablará de dos clases de corpus: los corpus grandes, de millones de palabras, que normalmente consultamos, por ejemplo, para explicar gramática o crear ejercicios. Y los corpus pequeños que podemos crear a partir de las producciones de nuestros alumnos con el fin de conocerlos mejor, poder crear contenido especializado para ellos o preparar un TFM sobre interlengua.

En particular, hablaremos de cómo decidir el tipo de corpus necesario para el estudio o la tarea que se tiene que realizar o de qué será representativo el corpus (por ejemplo, del habla general o del habla de los estudiantes italo hablantes de ELE del nivel B2). Después se verán, de manera muy resumida, los pasos necesarios para crear un corpus, desde su compilación hasta su anotación.

Este capítulo es, probablemente, el más técnico del libro, pero hay que pensar que el objetivo no es tanto saber crear un corpus de millones de palabras por nosotros mismos (ya hay lingüistas que se dedican a ello), como aprender, por una parte, la filosofía que hay detrás de los corpus para comprender los errores o incongruencias que puedan aparecer en las búsquedas, y, por otra, aplicar el conocimiento desarrollado por personas que consagran su vida a los grandes corpus a nuestros pequeños corpus de clase tomando prestadas todas las ideas de organización y anotación que sea posible.

4.1. TIPOS DE CORPUS

Al igual que se puede realizar una clasificación de los alumnos de un aula según diferentes criterios: como el nivel, la lengua materna, si prefieren destrezas orales o escritas, etc., también los corpus se pueden clasificar de acuerdo con diferentes criterios. En este epígrafe, clasificamos los corpus siguiendo muy de cerca los criterios usados por Torruella y Llisterri (1999), pero no sería la única clasificación posible.



Esquema 4. Tipología de corpus

Una de las primeras clasificaciones que se puede hacer es por el tipo de contenido del corpus. Así se puede diferenciar entre corpus escritos, orales o audiovisuales, en el caso de las lenguas de signos. En la misma línea, se pueden clasificar por soporte: texto (tanto los escritos como los orales transcritos), audio (orales con transcripción alineada) o audiovisual (en video con anotación de los gestos

y/o lenguas signadas). En ELE, ya sea en corpus creados por profesores a título particular o disponibles en línea, los **corpus escritos** podrían contener, por ejemplo, las redacciones de los estudiantes durante todo un año para ver su mejoría o redacciones de alumnos de A1 con diferente lengua materna recogidos durante 10 años. Un **corpus oral** contendría tareas orales, por ejemplo, los audios de una tarea de descripción de lámina, o diálogos entre dos estudiantes o entre el estudiante y el profesor. El corpus oral podría contener solo las transcripciones de esas grabaciones o el audio original. Por lo tanto, debemos tener en cuenta que, si queremos obtener errores de pronunciación, necesariamente tendremos que acudir a un corpus oral; en cambio si queremos obtener datos sobre cómo abren y cierran las cartas formales los estudiantes de ELE, el corpus tendrá que ser escrito.

Otro tipo de clasificación depende del tipo de lengua que contienen. Así, los corpus pueden contener lengua general o especializada (lenguaje jurídico, traducciones...), canónica (toda la obra de un autor), sincrónica (lengua actual) o diacrónica (diferentes etapas de la historia). En ELE, los **corpus generales** pueden servir para preparar clases o atender consultas de uso de una palabra por parte de los estudiantes, mientras que los **corpus de lengua especializada** sirven de apoyo para profesores de español con fines específicos. Existen corpus dedicados al turismo, al lenguaje jurídico, al lenguaje académico (Ainciburu, 2012). Hoy en día, están en auge los corpus de subtítulos y de localización, pero están más pensados para tareas de investigación.

Los **corpus canónicos** se usan sobre todo en estudios de literatura, por ejemplo, para caracterizar la lengua de un autor o para establecer la autoría de un texto comparándolo con toda la obra del autor que se cree que lo escribió. En principio, no se usan con alumnos de habla no nativa, a no ser en grupos de estudiantes universitarios que estén cursando asignaturas de lengua y cultura o literatura.

Por regla general, los **corpus diacrónicos** son aquellos que contienen datos históricos (como el CORDE). Aunque, de hecho, un corpus como Google NGrams, también está enfocado a la visión de la evolución en el tiempo de una palabra, así que hacer búsquedas diacrónicas con él es muy fácil. Además, hay que tener en cuenta que se puede usar un corpus diacrónico para hacer sincronía, se puede describir el uso de los pronombres de cortesía en el siglo XVI centrándose solo en eso y no en su evolución, y se estaría describiendo en sincronía. Por último, corpus que describiríamos como sincrónicos (como el SOL), pueden contener, por ejemplo, textos de los años setenta, y, por lo tanto, no serían adecuados sustantivos como *ordenador*, o *internet*, porque, aunque sean sincrónicos, no son actuales.

Según esta misma filosofía de clasificación, los corpus pueden ser de L1 o de L2 o lengua extranjera; estos últimos se llaman **corpus de aprendientes** o **corpus de errores** y, por su relevancia en nuestro campo, les vamos a dedicar todo un capítulo, el 6.

También podemos agrupar los corpus según el tipo de anotación que tienen. Así, un corpus sin anotación sería un **corpus crudo** (*raw*), y también tendríamos **corpus marcados** y **anotados** (veremos qué son en los apartados 4.4 y 4.5).

Un universo particular y especialmente interesante para el ELE es el de los **corpus multilingües**. En ellos podemos encontrar corpus comparables, paralelos, paralelos alineados... Veamos en qué se diferencian. Los **corpus comparables** contienen textos parecidos en dos o más lenguas: currículums, cartas de presentación (es decir, cartas formales), anuncios de alquiler y venta en inglés y en español, etc. Estos corpus tienen mucho potencial en nuestro campo porque en ellos contamos con ejemplos reales de textos que están muy convencionalizados para cada lengua. Por ejemplo, recursos que han usado nativos del español para poner un anuncio de venta que los estudiantes pueden usar como modelo y copiar adaptando los datos de los metros cuadrados o el número de habitaciones.

Otra clase de corpus multilingües son los **corpus paralelos**, que contienen el mismo texto traducido a otra lengua, por ejemplo, una novela y su traducción al español o una página web en su versión inglesa y española.

Si además el corpus paralelo está **alineado**, el texto traducido se muestra junto con el original para hacer más fácil la comparación. Este tipo de corpus son especialmente útiles para los estudiantes, y, de hecho, muchos estudiantes los usan aun cuando sus profesores de lengua extranjera no se los hayan mostrado nunca.

Vamos a ver cómo un estudiante podría llegar a un corpus paralelo alineado. Pongamos que está preparando una redacción y quiere decir que la ropa del personaje que está describiendo está muy gastada. En inglés, la palabra en la que estará pensando es *wear*. Si busca en Google “*wear in Spanish*”, el primer resultado será la traducción automática que dice *vestir*, pero el alumno sabe que ese no es el sentido de la palabra, que él no quiere la palabra *wear* en el sentido de *llevar puesto*, no quiere un verbo, que es lo que ha aprendido en clase. Si sigue bajando a los resultados de búsqueda, encontrará uno de Linguee que promete darle la palabra con ejemplos de uso. Lo que verá ahí (figura 3) serán ejemplos en inglés y en español. De este modo, el alumno puede buscar la frase en inglés que expresa lo que él quería decir y ver cómo se ha traducido en español.

Durable elastomeric backing protects the belt against environmental pollution as well as frictional wear if power is transmitted from the back of the belt. ↳ gates.com	[...] la correa de la contaminación ambiental y del desgaste causado por la fricción si se transmite la potencia por el dorso de la correa. ↳ gates.com
You will have to wear the bandages for at least 2 weeks after your initial treatment. ↳ turismoyasaludguale.com	Deberá utilizar las vendas por lo menos 2 semanas o hasta que su médico le indique. ↳ turismoyasaludguale.com
[...] must be designed and constructed in such a way as to prevent failure from fatigue and wear, taking due account of their intended use. ↳ europarl.europa.eu	[...] elevación se deben diseñar y fabricar de manera que se eviten los fallos debidos a la fatiga o al desgaste habida cuenta del uso previsto. ↳ europarl.europa.eu
Tires are subject to much less wear - longer service life. ↳ pottinger.fr	Las púas se desgastan mucho menos lo que se traduce en una vida útil más larga. ↳ pottinger.fr
A symptom of this wear is the spindle motor becoming noisier and read/write performance decreasing. ↳ storage.toshiba.eu	Un síntoma de este desgaste es que el motor de giro cada vez hace más ruido y disminuye el rendimiento de la lectura/escritura. ↳ storage.toshiba.eu
Edwards, a high school teacher, seems to always wear a smile on her face and have a thoughtful word of advice to give. ↳ levinechildrenshospital.org	Edwards, una profesora de escuela secundaria, siempre parece tener una sonrisa en su rostro y sabios consejos para ofrecer. ↳ espanol.levinechildrenshospital.org

Figura 3. Ejemplo de búsqueda de la palabra wear en inglés

Por último, según la cantidad (tamaño de muestra) y distribución de los contenidos, los corpus pueden ser de tres tipos. Pueden ser **corpus equilibrados**, que contienen la misma cantidad de muestras de cada periodo/autor, género, tipo o nacionalidad; **corpus monitores**, que van creciendo con el paso del tiempo para reflejar el habla actual; o **corpus oportunistas**, aquellos que se crean con un propósito concreto, por ejemplo, para ver el uso de la lengua de un grupo que ha seguido un programa lingüístico concreto. Este método de clasificación está ligado íntimamente con la noción de representatividad, por lo tanto, vamos a verlos con más detalle en el siguiente epígrafe.

4.2. REPRESENTATIVIDAD

Algo **representativo** es algo que refleja de manera fiel la realidad que quiere imitar. Así, un dibujo representativo de un perro será un dibujo en el que el perro tiene cuatro patas y se parece a un perro de verdad. Además, para que todo el mundo lo considere representativo es más probable que el dibujo sea de un pastor alemán que de un chihuahua. Esto último es la prototipicidad y, aunque es un concepto subyacente en los corpus (lo prototípico aparece más), no lo vamos a tratar de manera detallada. El concepto pertenece al ámbito de la lingüística cognitiva (Cuenca & Hilferty, 1999).

La primera pregunta que debemos hacernos cuando nos disponemos a recoger un corpus es: ¿*De qué queremos que sea representativo?* Si yo quiero hacer un corpus de “la lengua” y solo grabo adolescentes en el patio del instituto, obtendré una muestra **sesgada** en la que habrá, por ejemplo, muchos insultos (de hecho, eso es

algo que pasa en el corpus Val.Es.Co, un corpus más que recomendable para hacer reír a nuestros alumnos) (Briz Gómez et ál., 2002). Si, por el contrario, quiero una muestra **representativa** del español, es decir, **equilibrada**, necesito grabar a gente de todas las edades.

El concepto de **muestra** es clave o, más específicamente, el tamaño de la muestra. Cuando creamos un corpus, no tenemos acceso a toda la población, que en lingüística sería toda la lengua producida por todos los hablantes durante toda la historia. A lo único que tenemos acceso es a una muestra. Exactamente igual que en las encuestas del CIS cuando dicen que el 30 % de la población española hace algo, y los niños preguntan: *¿Y cómo lo saben? ¿Si a mí nadie me lo ha preguntado!* En lingüística, decimos que los hablantes usan el subjuntivo detrás de *no creo que*, pero no se lo hemos preguntado a todo el mundo, tenemos **muestras** del habla.

Solo hay un tipo de corpus en el que no tenemos una muestra, sino toda la lengua: se trata de los corpus **canónicos**, que contienen toda la obra de un autor. En este caso, como tenemos toda la obra del autor, en términos estadísticos, no tenemos una muestra de la población, sino la población al completo. Estos corpus son perfectamente representativos, aunque poco útiles para nosotros.

El resto de los corpus, para que sean representativos, tienen que estar bien diseñados. Antes de recoger los datos, hay que pensar cuántos datos queremos recoger de cada estamento o de cada grupo. Así, si queremos hacer un corpus de novela, nos plantearemos qué tanto por ciento de novelas de cada género se incluirán. ¿Incluiremos el mismo porcentaje de cada género? ¿O quizá preferimos incluir más novelas históricas porque se publican más que, por ejemplo, de ciencia ficción?

Lo mismo se aplica a cualquier tipo de corpus. Tomando el ejemplo de un corpus de ELE, ¿cuántos alumnos tenemos que incluir? ¿Tenemos que incluir los mismos alumnos de A1 que de C2? La respuesta a esta pregunta no es fácil y depende del objetivo de nuestro trabajo. Es cierto que en el mundo hay más estudiantes de A1 que de C2. Por lo tanto, si queremos una muestra representativa de los alumnos de español en el mundo escogeremos más de A1. No obstante, si el corpus es para realizar un análisis, nunca vamos a mezclar los datos de A1 con C2, porque los tipos de errores que cometen en cada nivel, las estructuras que usan, todo es demasiado diferente, así que haremos dos análisis y después los compararemos. Como el resultado final será una comparación entre grupos (grupo de A1 y de C2) para que esté equilibrado, necesitaremos tener el mismo número de alumnos en los dos grupos.

En las clasificaciones del apartado anterior hemos visto que existían corpus llamados equilibrados (*balanced*), monitor y oportunistas. Visto en términos de representatividad, un **corpus equilibrado** sería el más representativo y el que tiene un diseño más riguroso. Este tipo de corpus es útil en términos de calidad cuando se está creando un corpus muy grande para una editorial que creará un diccionario, pero no siempre más diseño quiere decir mejor.

Un **corpus monitor** tiene un diseño menos definido porque, según va creciendo para reflejar el habla de cada momento, las proporciones de cada tipo de texto pueden ir cambiando. Precisamente por eso, es un corpus que refleja el habla actual de manera más fiel que un corpus general que tenga un diseño más equilibrado. Este tipo de corpus es al que acudiremos para hacer consultas de lengua, para saber cómo es el español hablado por nativos, para explicar gramática a nuestros alumnos, para tomar ejemplos de habla real o para dejar que los mismos alumnos hagan ejercicios, pero no es el tipo de corpus que un profesor va a construir, a no ser que haga un cambio de carrera, deje las clases (o la formación) y entre a trabajar como lingüista computacional. En cambio, sí que son útiles para usar, y, para ello, es importante que saber sus criterios de diseño y ser conscientes de 1) qué tipo de lengua tiene, 2) cuántos millones de palabras contiene o, 3) como en el caso del corpus SOL que vimos en los ejercicios del capítulo 2, qué años cubre.



Esquema 5. Tipos de corpus ordenados según su representatividad

Probablemente, ya sea para realizar un trabajo académico, como creador de materiales o como profesor, para tener una idea de los errores que comete un grupo de alumnos, se va a crear un **corpus oportunista** porque no interesan todos los alumnos de ELE del mundo. Los corpus oportunistas, pese a la connotación negativa de la palabra, no son malos, simplemente, están diseñados para responder una pregunta concreta, y, por lo tanto, solo se pueden usar para el trabajo para el que fueron diseñados, mientras que un corpus general de la lengua puede servir para dar respuesta a todo tipo de trabajos.

Cuando tenemos un objetivo concreto, necesitamos un corpus específico y eso significa crear un corpus oportunista. Por ejemplo, en el caso de un estudio en el que se quieren analizar errores cometidos en el subjuntivo en alumnos de nivel C

de origen italo hablante, necesitamos justamente eso y nada más. Por lo tanto, el corpus que crearemos estará sesgado, porque no incluirá alumnos de diferentes L1, pero es lo que nos interesa para el objetivo de nuestro trabajo.

En el caso de trabajar con un grupo de clase, nos interesarán todas las redacciones que han producido nuestros alumnos durante el año, quizás en una tarea concreta, para ver su tipología de errores. Ese corpus será representativo solo de nuestra clase de ese año y no de todos los grupos que hemos tenido, ya que un grupo puede tener muchos problemas con el uso del imperfecto y el indefinido, y al año siguiente ninguno porque los estudiantes provienen de una lengua materna diferente. En cualquier caso esos son los datos que nos interesan en ese momento.

La diferencia tiene que ver con qué se quiere representar: un corpus oportunista es representativo de una sección muy pequeña de la lengua, mientras que un corpus equilibrado tiene como objetivo representar **la lengua**, si es que eso es posible.

4.3. COMPILACIÓN

Una vez decidido el tipo de corpus que se creará y su tamaño, el primer paso para crearlo es recoger (recopilar) los materiales que formarán parte de este.

En el caso de los grandes corpus textuales (equilibrados o monitor), esa compilación suele consistir en ir a una biblioteca, recoger una serie de libros y revistas y escanearlos con un reconocedor óptico de caracteres (OCR) que permitirá su digitalización. Este proceso se realiza todavía hoy para la digitalización de libros impresos. Es lo que hace Google Books, pero también corpus de referencia que están digitalizando obras publicadas en el siglo xx, como el CORPES.

Sin embargo, hoy en día, muchos corpus nacen digitalizados; son textos disponibles en la web, que no tienen un formato impreso. Por ejemplo, son muy abundantes los corpus consistentes en textos de la Wikipedia, de Twitter o de las opiniones que dejan los usuarios en Amazon (Fornaciari & Poesio, 2014; McCreadie et ál., 2012; Reese et ál., 2010). En este caso, la compilación es diferente, ya que no es necesario digitalizar los textos. Para recoger este tipo de corpus se crea un pequeño programa informático que se encarga de recoger el texto que se va generando en la web (por ejemplo, en Twitter) y lo guarda en ficheros txt o en una base de datos (*web scraping*). Desde luego, como profesores de ELE, no es necesario que nosotros lleguemos a esto (a no ser que tengamos interés); los lingüistas computacionales se encargan de recoger esos textos y ofrecérmolos con una interfaz de búsqueda bonita y fácil de usar.

En el caso de los corpus orales, se pueden usar grabaciones recogidas de televisión, entrevistas, o provenientes de contextos de interacción específicos como por ejemplo las preguntas en una oficina de venta de billetes de tren. En el caso de los corpus de segunda lengua, esas grabaciones serán de tareas de clase.

Todos los corpus compilados que se han explicado hasta aquí son tipos de corpus profesionales, que se crean para su explotación por parte de una empresa, y son, por tanto, corpus de gran envergadura. Aunque, una vez más, es poco probable que alguien que se dedica al mundo del ELE tenga que compilar un corpus tan grande.

En general, el profesor de ELE tendrá bastante con explotar los corpus que hayan recogido otras personas, pero también es posible que necesite compilar corpus propios. En ese caso, será un corpus de aprendientes de ELE, es decir, específico, y si se hace para una investigación en concreto (por ejemplo, para el trabajo final de máster) será, además, un corpus oportunista. Este tipo de corpus, dado su tamaño, requerirán de técnicas mucho más artesanas para su recopilación.

Básicamente, el corpus podrá consistir en:

1. las producciones textuales de nuestros alumnos durante un periodo determinado, por ejemplo, sus redacciones o los correos electrónicos que nos escriban. Siempre estarán digitalizados, es decir, si nuestros alumnos nos entregan las redacciones escritas a mano, tendremos que copiarlas en el ordenador (con sus errores incluidos, así que habrá que desconectar el autocorrector), para después poderlas procesar de manera automática;
2. en las producciones orales de nuestros alumnos grabadas digitalmente y, posteriormente, transcritas.

4.4. MARCAJE O CODIFICACIÓN DE CORPUS (MARK-UP)

Imaginemos que durante años recopilamos redacciones de nuestros estudiantes para convertirlas en un corpus. Cuando recuperamos todos esos archivos, pongamos 200, nos damos cuenta de que el nombre del archivo es el que tenía cuando los alumnos nos lo enviaron por correo electrónico: a veces su nombre, a veces el nombre de la tarea, a veces incluyen el grupo, a veces el nombre del archivo es simplemente "español.doc" y el alumno ni siquiera incluyó su nombre dentro de la redacción (¡qué escenario tan habitual!).

Hemos recopilado el corpus y lo tenemos digitalizado, pero no tenemos ninguna pista sobre cómo ordenarlo. Para evitar eso, en nuestros corpus oportunistas,

normalmente ponemos un nombre de archivo que sea útil para la clasificación, algo como “2020_septiembre_descripción_L1inglés_B1.docx”. Ese tipo de nombres nos ayudan a identificar el contenido del fichero.

Pues eso, pero un poco más complejo, es el **marcaje** o codificación de corpus (*mark-up*). Llevando la terminología de corpus al ámbito de la lingüística más tradicional, podríamos decir que el marcaje explicita datos del contexto (autor, lugar, fecha...). Los datos que se incluyan en el marcaje dependerán, de nuevo, del objetivo del corpus. Por ejemplo, si estamos haciendo un estudio experimental para trabajar la pronunciación con dos grupos de alumnos y en uno hemos trabajado con método verbo-tonal (Billières, 2015) y, en el otro, con método audiolingüe, incluiremos esa información en el marcaje del corpus. Así, cuando tengamos que analizar los resultados, comprobar qué grupo ha aprendido más, nos será mucho más fácil porque ya tendremos los datos clasificados. Si, en cambio, estamos creando un corpus de novelas, incluiremos la fecha de publicación de la novela, los años que tenía el autor cuando la publicó o su profesión.

Si los usuarios del corpus vamos a ser nosotros mismos, crearemos un sistema de marcaje que nos sea útil para nuestro propósito y que para nosotros sea natural, que esté en línea con nuestra manera de pensar. Pero ¿qué pasa en los grandes corpus en los que hay muchos compiladores que trabajan a la vez y muchos usuarios potenciales? En esos casos, el marcaje tiene que ser una solución de consenso, algo que sea útil para la mayor cantidad posible de usuarios y, además, algo que todas las personas que compilan y marcan el corpus hagan de la misma manera. Por eso, a lo largo de la historia, ha habido grandes iniciativas en las que se han propuesto maneras sistemáticas de anotar un corpus.

Para ver una revisión completa de sistemas de codificación se puede consultar Hardie (2014). Algunas de las iniciativas más populares han sido la *Text Encoding Initiative* (TEI) (Ide & Véronis, 1995), COCOA (McEnery & Wilson, 1996) o *Standard Generalised Markup Language* (SGML), si bien ahora se prefieren métodos basados en etiquetas XML que son también las usadas actualmente en biblioteconomía.

En la figura 4, se puede ver en qué consisten “las tripas” del marcaje. Sin embargo, como usuarios finales de corpus, no necesitamos ver esas tripas. Los corpus con los que vamos a trabajar como usuarios nos permiten hacer búsquedas basadas en un autor o en un periodo de tiempo gracias a que están marcadas de esa manera, pero en la visualización final el marcaje no aparece.

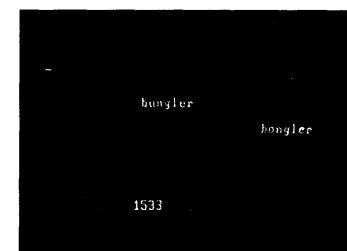


Figura 4. Marcaje del Oxford English Dictionary (fuente: Wikimedia Commons)

Saber cómo funciona el marcaje nos sirve para entender cómo son posibles las búsquedas y por qué, por ejemplo, Google NGrams nos puede devolver como propia del año 2007 la forma medieval *mientre*. ¿Cómo puede pasar eso? Pues porque muchas ediciones modernas de textos antiguos, por ejemplo, publicadas en los 2000, están marcadas como primera edición del texto (lo cual es cierto) y, entre ellas, en el año 2007 hubo un texto que empezaba así: “De los sos ojos tan fuerte mientre lorando” o que, simplemente, citaba este archiconocido inicio del *Cantar de Mio Cid*.

La segunda cosa que podemos aprender del marcaje de textos profesional es la estrategia de marcaje que se debe tener para los corpus de clase. Para nosotros, el marcaje no tiene por qué estar en un archivo XML. Puede ser simplemente un nombre de archivo claro, que nos permita recuperar toda la información necesaria del autor de la redacción. La información que debería contener aplicada al ELE de manera potencial (no siempre hace falta tener en cuenta tantas variables) podría ser:

1. **Identificador del alumno:** puede ser el nombre o un código para preservar su anonimato.
2. **Nivel del alumno/Grupo:** de A1 a C2, también se puede marcar como principiante/avanzado, etc. Si hay dos grupos se pueden marcar como “el de los martes a las 18”: MA18, y “el de los miércoles a las 4”: MI16. Si se está usando un método diferente con cada uno de ellos, indicar el método que se usa: “audiolingüe” o “verbo-tonal”.
3. **Tarea:** si es una tarea oral o escrita. Si es una tarea de descripción o de narración. Si es una carta formal o informal. Si es un diálogo o un monólogo. Todas estas facetas de la tarea pueden ser interesantes para después facilitar el análisis: por ejemplo, si queremos ver usos del indefinido, buscaremos narrativas (contar una anécdota o una biografía), ya que el indefinido es un tiempo que tiene una frecuencia alta con ese tipo de tareas.
4. **Tiempo:** Si vamos a recoger muestras durante todo el curso, puede ser útil marcar qué producciones son de septiembre y cuáles de junio, o simplemente numerarlas por orden cronológico.
5. **L1 del alumno:** preferentemente la L1 real, es decir, si un alumno es bilingüe, por ejemplo, sinohablante, pero conoce también otra lengua del país de origen, anotaremos la lengua que sea lengua materna.
6. **Edad:** edad y no fecha de nacimiento, porque lo que nos interesa es si era un alumno adolescente, un niño o un anciano que pudiera tener un deterioro cognitivo que afecte a su memorización del léxico.
7. **Edad de inicio de adquisición:** si es un aprendiz adulto, su adquisición empezó en la niñez o en la adolescencia. Esta variable es especialmente importante si se hacen estudios de pronunciación.
8. **Periodos de inmersión:** si ha vivido en un país hispanohablante, si es que sí, durante cuánto tiempo.
9. **Sexo y género:** Por regla general, si el objetivo es hacer un estudio de modelos de aprendizaje, nos interesará contemplar como variable el género (el papel social que asume el estudiante), pero, para corpus orales, puede ser interesante contar con el sexo biológico codificado para poder aplicar correcciones automáticas por el timbre de voz (las mujeres tienen la voz más aguda que los hombres).

Por supuesto, no es necesario que todos los corpus contengan tanta información de los estudiantes o de las tareas; como siempre, dependerá del uso que queramos hacer de los datos.

4.5. ANOTACIÓN

Si el marcaje daba información del contexto en el que se había producido el mensaje, la **anotación** analiza la lengua del mensaje en sí. Esto posibilita que, en una búsqueda en corpus, se pueda pedir a la interfaz “devuélveme los casos en los que *médica* está usado como nombre”, tal y como hemos hecho en los ejercicios. Esto es porque detrás de ese corpus, en sus tripas, como decíamos para el marcaje, existe una anotación morfológica realizada con la ayuda de un anotador automático, lo que se llama un POS-*tagger*, y una sintáctica realizada con la ayuda de un anotador sintáctico automático (*parser*). Estas dos anotaciones son las más clásicas del mundo del corpus, pero no las únicas posibles.

Los corpus se pueden anotar en todos los niveles gramaticales: fonético-fonológico (sonidos), prosódico (melodía de la frase), morfológico (clase de palabra –artículo, nombre, adjetivo, preposición...–), sintáctico (función en la frase –sujeto, objeto, complemento...–), semántico (significado). Además, se pueden hacer anotaciones especializadas para cada ámbito. Por ejemplo, en lingüística forense se anotan los corpus de entrevistas policiales (interrogatorios a sospechosos) con información sobre si un enunciado es “verdad”, “mentira”, “probablemente verdad”, “probablemente mentira” o “no comprobable”. De esta manera, después se puede estudiar las características del discurso falso e incluso entrenar sistemas para detectarlas de manera automática (Fornaciari & Poesio, 2012).

También en los corpus de segunda lengua existe una anotación especializada, la anotación de errores. Es cierto que en un corpus de ELE puede ser útil contar con una anotación morfológica, por ejemplo, para ver el número de conectores de tiempo (adverbios de tiempo) que usan los alumnos. De esta manera, podríamos estudiar de manera cuantitativa, basándonos en datos y con porcentajes una característica de la interlengua del español L2, pero habitualmente los corpus de L2 están enfocados a la anotación del error. Anotar errores es básicamente pasar las correcciones de las producciones de los alumnos a un formato digital. Pero ¿cómo se pueden marcar los errores?

La clasificación de los errores depende de cada corpus. De manera general, los corpus anotan el nivel del error (error de pronunciación, morfológico, sintáctico, léxico...) y después hacen subcategorías con los errores más comunes o interesantes. Por ejemplo, error de concordancia de número (*la gente *eran*), error de concordancia de género (**el radio*), error de tiempo verbal (*ayer *voy*), error de persona verbal (*ellos *vamos*), confusión entre *ser* y *estar* (**soy en casa*)...

De nuevo, para la anotación de nuestros corpus (sobre todo si es para un trabajo concreto) quizá solo nos interese anotar un tipo de error dependiendo de lo que estemos estudiando.

La anotación de los corpus lleva mucho tiempo y, si no estamos creando un corpus para que esté disponible en línea para un gran número de usuarios, no tiene sentido que le dediquemos horas a un trabajo que después no vamos a explotar. Para dar una idea de la cantidad de trabajo que supone, el tiempo necesario para transcribir un minuto de grabación es de una hora, así que quizá nos merece la pena hacer un corpus más oportunista y transcribir tan solo aquellos pasajes que sean interesantes para nuestro objeto de estudio.

4.6. PARA MÁS INFORMACIÓN

Torruella, J., & Llisterra, J. (1999). Diseño de corpus textuales y orales. Filología e informática. Nuevas tecnologías en los estudios filológicos, 45-77.

4.7. ACTIVIDADES

4.7.1. Busca la información de diseño del corpus CREA, disponible aquí (<https://www.rae.es/recursos/banco-de-datos/crea-escrito>). ¿Crees que es un corpus equilibrado respecto al origen de las obras? ¿Por qué?

4.7.2. En el mundo del ELE, ¿un corpus de errores de alumnos de L1 de todas las lenguas del mundo estaría equilibrado? ¿Sería útil?

4.7.3. El corpus CLUVI, recopilado por la Universidad de Vigo, es el corpus más completo en español de lenguajes especializados y traducción. Es uno de los corpus que se puede usar en español con fines específicos. Para ver algunas de sus funcionalidades ve a la web: <http://sli.uvigo.es/CLUVI/index.php#>. En la parte superior izquierda, PESCUDAR TEXTOS, encontrarás el acceso a los corpus organizados por tipo lengua especializada y lenguas:

1. En el corpus de terminología legal, busca como se traduce el “fallo” jurídico de español a gallego.
2. Busca la traducción de *totally* al español y de *completamente* al inglés en textos literarios. ¿Qué puedes decir sobre cómo se traducen esos adverbios?
3. ¿Cuál es la traducción de habitual de *ignore* al español en textos literarios? ¿Y en subtítulos?

4.7.4. Hace unos años, el corpus de libros publicados de Google Books se habilitó como corpus a través de una interfaz de búsqueda (<https://books.google.com/ngrams>). Como explicaban en la charla TED recomendada en el capítulo 2, este corpus es uno de los más visuales para ver cambios culturales en español. En el capítulo 8 veremos cómo llevarlo a clase de ELE, pero de momento, vamos a explorarlo. En la parte superior, escoge el corpus “Spanish 2012” y busca en esa interfaz de manera conjunta la palabra *médico* y *médica* y observa su frecuencia a través del tiempo.

1. ¿A qué crees que se deben las ocurrencias de *médica* anteriores al siglo xx?
2. Asegúrate de que estás buscando en la versión de 2012 (está lematizada, la versión de 2009, no). Ahora realiza la siguiente búsqueda: “médico_NOUN, médico_ADJ, médica_ADJ, médica_NOUN”. ¿Podrías explicar ahora qué ha pasado? Para ver todas las formas de búsqueda que permite el corpus consulta <https://books.google.com/ngrams/info>.

5

APLICACIONES DE LOS CORPUS EN LA ACTUALIDAD

El uso que nosotros le daremos a los corpus, la didáctica de la lengua, no es el más habitual ni el que mueve más dinero. El desarrollo de corpus anotados en la actualidad está ligado con el desarrollo de las herramientas para el procesamiento de lenguaje natural. Es decir, en el mundo empresarial, los corpus no se están usando para enseñar lenguas a humanos, sino a máquinas. Es gracias a los corpus que asistentes virtuales como Alexa, Cortana o Siri pueden responder a nuestras preguntas y nos entienden cuando les hablamos. No obstante, este uso de los corpus queda un poco alejado del objetivo de este manual, así que, en este capítulo, vamos a ver algunas aplicaciones de los corpus que quedan más cerca del quehacer del profesional de la lengua, aunque no sea estrictamente profesor de ELE. Específicamente, hablamos de su aplicación a la lexicografía, a la investigación lingüística, a la lingüística forense y a la lingüística clínica.

5.1. LEXICOGRAFÍA

Tal y como hemos visto en la sección de historia, uno de los primeros usos de los corpus fue la creación de diccionarios, es decir, la lexicografía (Rojo, 2009).

Si pensamos en la historia de los diccionarios, estos llevan siglos incluyendo en sus acepciones ejemplos de uso y, en algunos casos, de uso real de la lengua. En el caso del español, el primer diccionario de la Academia, el llamado *de Autoridades*, se llamó así porque incluía ejemplos de uso de autores considerados clásicos. Por ejemplo, bajo el lema *patata* se recoge el ejemplo de Quevedo ilustrado en (1) con el estilo políticamente incorrecto que le caracteriza.

1. Quev. Orland. Cant. I. Los Andaluces de valientes feos, cargados de patátas y cecêos.

Hoy en día, los diccionarios siguen incluyendo ejemplos, pero, cada vez más, estos están extraídos de corpus, por ejemplo, de noticias. Aunque este no es su único uso.

Para empezar, tal y como pasaba en el COBUILD, los corpus se usan para seleccionar el léxico que se incluye en los diccionarios. En algunos casos, los diccionarios aspiran a incluir todo el lexicón de una lengua, pero, en otros, se incluyen solo las **palabras más frecuentes**. La determinación de cuáles son las palabras más frecuentes se hace mediante corpus.

Muchos diccionarios, además del significado de la palabra, incluyen alguno de sus patrones colocacionales. Por ejemplo, para los verbos que rigen preposición se incluye esa preposición, como en *llenar de*. En el caso de los diccionarios combinatorios, se incluyen el lema y las palabras junto a las que aparece más frecuentemente, como *llenar a medias* o *llenar de alegría*.

Sabemos que existen corpus enriquecidos con anotaciones que permiten extraer de manera automática información como las categorías gramaticales. La existencia de estos corpus permite que algunos diccionarios incluyan información sobre cómo se usa la palabra, es decir, ¿se usa el adjetivo *conocido* como nombre (*me encontré a un conocido*)? ¿Es más frecuente el uso de *joven* como nombre (*el joven*) o como adjetivo (*el hombre joven*)?

En el caso de que un corpus contenga información semántica, esta también se puede aprovechar para la realización de diccionarios. Por ejemplo, se pueden extraer todas las palabras de un campo semántico o confeccionar un diccionario de hiperónimos y sus hipónimos, o agrupar palabras por polisemia o homonimia.

Por último, el enriquecimiento constante de los corpus (monitor) permite que se puedan incluir rápidamente los extranjerismos y monitorizar como se está adaptando un préstamo de reciente incorporación, una información que es fundamental en un diccionario de uso. Por ejemplo, ¿es más habitual, *marketing*, *márketing* o *márquetin*? ¿Cuál de estas tres opciones debería aceptar la Academia?

5.2. INVESTIGACIÓN LINGÜÍSTICA

Los corpus también son una herramienta útil para la investigación básica en lingüística. Con ellos se puede describir un fenómeno lingüístico o una lengua completa. Se pueden hacer descripciones generales de la lengua o de una variedad concreta, ya sea una variedad dialectal, sociolectal o un registro.

Tal y como pasaba con los diccionarios, los corpus se han usado para crear gramáticas desde hace mucho tiempo. En paralelo a la creación del COBUILD, se creó una gramática basada en el mismo corpus: se trata de *A Comprehensive Grammar of the English Language* (Quirk et ál., 1985). Otra gramática clásica realizada usando

un corpus como base es *Longman Grammar of Spoken and Written English* (LGSWE, Biber et ál. 1999).

Las ventajas del uso de corpus en la investigación lingüística radican en que se pueden comprobar las hipótesis con datos, por lo tanto, hacen que el estudio empírico del lenguaje sea posible y, además, permiten la cuantificación de los fenómenos. Así, afirmaciones basadas en las creencias de los lingüistas, como que las mujeres hablan más que los hombres, (específicamente. Liberman afirmó en 2006 que los hombres usan exactamente 7000 palabras al día mientras que las mujeres usan 20 000 (Liberman, 2006)), se ven refutadas por los datos que nos muestran que tanto hombres como mujeres dicen unas 15 000 palabras al día (Mehl et ál., 2007).

El uso de corpus facilita estudios que no se podrían realizar de modo introspectivo como, por ejemplo, las diferencias entre el uso oral y el escrito del lenguaje, usos particulares del discurso oral, estudios de dialectología... Como se puede observar las posibilidades de los estudios basados en corpus son infinitos y, de hecho, cualquier investigación lingüística que se precie en la actualidad hará uso de corpus.

5.3. ESTUDIOS DE TRADUCCIÓN

Los estudios de traducción se caracterizan por el uso (junto con los de lingüística contrastiva), de los corpus multilingües o bilingües, y paralelos (alineados o no).

Estos corpus se usan en los estudios de traducción para avanzar en sus teorías sobre los rasgos de la lengua traducida, el llamado *traducciónés* (Corpas Pastor, 2008). Además, a partir de estos corpus se realizan comparaciones de estructuras lingüísticas en dos lenguas y se cuantifican. Vamos a ver un ejemplo entre español e inglés con las palabras *completamente* y *absolutamente*. A partir de corpus, se puede estudiar qué tanto por ciento de las veces *absolutely* se traduce por *absolutamente* y en qué casos por *completamente* y al revés, cuando *completely* se traduce por *completamente*. El resultado de este estudio permite mostrar que la distribución es diferente en inglés que en castellano, es decir, *absolutely*, no siempre se traduce por *absolutamente* y *completely* no siempre se traduce por *completamente*.

Otro ejemplo interesante es el de los verbos *causar* y *cause*. A partir del análisis del contexto en el que aparecen, se puede observar que, en español, *causar* se puede usar con predicados con una carga semántica positiva o negativa, es decir, se pueden causar cosas buenas y malas, pero en inglés no. De las apariciones de

causar en castellano solo aquellas que causan cosas negativas se traducen al inglés con el verbo *cause*. Por lo tanto, los contextos en los que aparece *causar* en español no son totalmente equiparables con los de *cause*; no es una traducción perfecta.

Esto se aplica después a la enseñanza, mostrando a los estudiantes de traducción los contextos de traducción de cada palabra para que puedan refinar sus traducciones. Además, estos corpus se usan para entrenar sistemas de traducción automática. Es decir, los sistemas informáticos aprenden que *house* se traduce como *casa* a partir de corpus paralelos donde se ha traducido de esa manera.

5.4. LINGÜÍSTICA FORENSE

La lingüística forense saca partido de los corpus en dos grandes campos de estudio. El primero es el estudio del lenguaje jurídico y su simplificación para hacerlo más accesible a los ciudadanos. El segundo (que más interés suscita) es el uso de muestras de lengua como prueba en un juicio.

Para el primero, se usan corpus que contienen lenguaje jurídico, como sentencias o actas, y se analiza su uso de la lengua para proponer versiones simplificadas que hagan las leyes entendibles para el ciudadano medio.

Sobre el segundo, se pueden crear y usar corpus de muchos tipos dependiendo de la tarea que haya que realizar (Coulthard, 2013). Por ejemplo, en casos de detección de autoría y plagio, se comparan dos textos o dos grabaciones (por lo tanto, el corpus está formado por dos muestras) para determinar si pueden pertenecer a la misma persona. Si la tarea que hay que realizar es la determinación de origen del locutor, se recoge un corpus de la persona investigada y se usan, como corpus de referencia, corpus de diferentes dialectos para comprobar a qué dialecto se parece más el modo de hablar o de escribir (sobre todo en el caso de crímenes cibernéticos) del investigado.

5.5. LINGÜÍSTICA CLÍNICA

La lingüística clínica se encarga de estudiar las características del habla patológica y en su vertiente más aplicada evaluar y rehabilitar pacientes (sus profesionales son logopedas y fonoaudiólogos) (Crystal, 2013). Para hacerlo, la lingüística clínica puede hacer uso de dos tipos de corpus.

En primer lugar, los corpus generales de la lengua. Este tipo de corpus se pueden usar para crear textos fonéticamente equilibrados (que contienen todos los sonidos de la lengua en la misma proporción que un texto real) y para realizar

listas de vocabulario básico o de alta frecuencia. En los dos casos, el resultado es un texto o una lista de vocabulario que usan los logopedas para evaluar o rehabilitar el lenguaje.

El segundo son corpus de hablas patológicas donde se ha grabado y transcrito a pacientes. Este tipo de corpus son los que se usan para describir el habla afectada y así poder determinar qué ámbitos hay que rehabilitar o para describir el habla de cierta patología o condición. Por ejemplo, la prosodia en personas con síndrome de Asperger, o incluso determinar en qué estadio se encuentra una enfermedad degenerativa, dependiendo del tipo o la gravedad del error.

6

CORPUS PARA INVESTIGAR SOBRE EL ESPAÑOL LENGUA EXTRANJERA

Hasta este momento, hemos visto una introducción a los corpus y a sus tipos, hemos aprendido la terminología del campo y hemos visto algunas de las formas en las que se puede explotar un corpus (hacer búsquedas). Pero en todo este tiempo, aunque haya habido ejemplos de ELE, no se han abordado tareas específicas de ELE. A partir de ahora, vamos a hablar de cómo explotar un corpus, específicamente en el campo del ELE. En primer lugar, en este capítulo, vamos a tratar cómo hacer investigaciones de ELE basadas en corpus y, en capítulos posteriores, veremos cómo usar corpus para crear materiales, cómo llevarlos a clase para que los alumnos los puedan usar en actividades y cómo crear corpus para corregir mejor.

Empezamos con la investigación. La investigación en el mundo del ELE gira alrededor de crear prácticas, didácticas más efectivas (Brown & Rodgers, 2002). Para crear esas prácticas en lo primero que hay que fijarse (o más bien lo segundo, después de un análisis contrastivo entre la lengua meta y la lengua de origen del estudiante) es en los errores que cometen los alumnos. Por eso, para poder ver una gran cantidad de errores cometidos por alumnos de ELE, existen los corpus de aprendientes. Este tema recoge las posibilidades de la investigación en ELE, define y explica los corpus de aprendientes y hace una revisión de los corpus de aprendientes de español disponibles en línea.

6.1. ¿QUÉ SE PUEDE INVESTIGAR EN ELE?

La mayor parte de las investigaciones sobre ELE se realizan desde el mundo universitario, bien sea por parte de profesores especializados, doctorandos o alumnos que preparan su trabajo final de máster (TFM) o incluso su trabajo final de grado (TFG). En todos estos casos, para preparar sus trabajos o artículos, los investigadores pueden contar con corpus ya recopilados o preparar un corpus propio con un grupo concreto de estudiantes en el que están interesados. Por ejemplo, un grupo de sinohablantes de nivel B1 y otro de nivel C2 para comparar sus resultados.

Pero eso no significa que los profesores de ELE no investiguen. En realidad, ¡lo hacen constantemente! La diferencia es que cuando un profesor de ELE realiza una investigación, normalmente la hace sobre el grupo con el que trabaja en esos momentos. Es decir, creará su propio minicorpus (un corpus oportunista), que contendrá toda la información que tiene de su grupo de alumnos de determinado año junto con, por ejemplo, sus redacciones. Así, el profesor conocerá a sus alumnos, sus dificultades y sabrá si en su próxima clase tiene que volver a repasar los usos del subjuntivo o el léxico de la familia. Esto es también investigación, pero con algunas diferencias respecto a la investigación de las universidades.

La mayoría de los profesores no crean un corpus propiamente dicho (no lo necesitan), sino una base de datos, ya que, como hemos repetido varias veces, un corpus necesariamente tiene que contener las redacciones digitalizadas y muchas veces el profesor las tendrá en papel. Pero quizá, a partir de ellas, haya extraído una lista de errores frecuentes que haya colocado en un Excel o simplemente se haya hecho una lista de cosas que no han quedado claras. Esos apuntes-resumen que tomamos los profesores al corregir una tarea no serían un corpus, sino más bien una base de datos. Crear un corpus digitalizado y completo de las producciones de los alumnos llevaría un tiempo que, si no se va a explotar el corpus más tarde (se va a poner en línea o se va a compartir por toda la escuela, por ejemplo), no merece la pena. Sin embargo, saber la metodología de los corpus de investigación hace que las bases de datos estén organizadas de manera más eficaz, porque se pueden aplicar marcajes o anotaciones que se han usado en los grandes corpus de ELE.

Seas profesor, formador, investigador o estés haciendo un TFM (o trabajo final de asignatura) de investigación, el objetivo final es determinar y/o mejorar la eficiencia del aprendizaje. Y para explicar esa eficiencia, se pueden tener en cuenta varios factores: 1) factores cognitivos (inteligencia, memoria...) y factores personales del aprendizaje (motivación, autoimagen...) y 2) factores relacionados con las diferencias de los sistemas lingüísticos de los alumnos.

Para estudiar las primeras, no se usan corpus propiamente dichos, sino entrevistas personales con datos sociolingüísticos asociados que nos muestran las actitudes de los alumnos a través de preguntas directas o de escenarios comunicativos controlados. Esta línea de investigación, que se ha trabajado desde el construccionismo social (Williams & Burden, 1999), no la trataremos en este libro.

En cuanto a las dificultades que supone la lengua meta dependiendo de la L1 de los alumnos, existen tres líneas de investigación principales, 1) el análisis contrastivo,

2) el análisis de errores y 3) el análisis de interlengua. Todas ellas, al tener como objeto de estudio la lengua, se estudian a partir de corpus.

El análisis contrastivo supone hacer una comparación de los sistemas de dos lenguas para observar sus diferencias y así prever los casos de transferencia positiva y negativa que tendrán los aprendientes. Por ejemplo, si pensamos en la pronunciación, los francófonos tendrán problemas para pronunciar la erre [r] pero no la eñe [ɲ], ya que la primera no existe en sus lenguas, pero la segunda sí. En cambio, los anglófonos sí tendrán problemas con el sonido nasal palatal [ɲ], ya que no existe en inglés. Si pasamos a ver un ejemplo en el léxico, los anglófonos pueden tener problemas para usar correctamente el verbo *creer* para expresar opinión y harán, en su lugar, un uso demasiado habitual de *pensar*. Es decir, usarán la construcción “Yo pienso que...”, en ocasiones en que un nativo hubiera preferido “Yo creo que...”.

Para hacer un análisis contrastivo, se pueden usar corpus generales o de referencia de las dos lenguas a comparar. Si volvemos al ejemplo de pensar/creer, podríamos consultar el CORPES para el español y el BYU para el inglés y obtener las concordancias de los verbos *creer* y *believe* para comprobar si tienen los mismos usos. Pero esta no es la manera más directa de hacerlo. La manera más rápida de acometer un análisis contrastivo es a partir de corpus paralelos, de los que ya hemos hablado en capítulos anteriores. A partir de estos, podemos ver como se han traducido las piezas léxicas en la otra lengua. Siguiendo con el ejemplo de *creer* que hemos usado arriba, una búsqueda en Linguee (<https://www.linguee.es/espanol-ingles/traduccion/creer.html>) revela que *creer* en el sentido de “pensar u opinar algo” se traduce por *think* y no por *believe*. Y ahí tendríamos la explicación para la transferencia negativa que observábamos arriba que causa que los anglófonos usen el verbo *pensar* (“Yo pienso que”) en momentos en que, en español, sería más adecuado usar *creo que*. *Creer*, en su lengua, no se usa para expresar opinión.

La segunda línea de investigación es el análisis de errores. Este consiste en describir las producciones desviadas de los alumnos. El análisis de errores puede ser de cualquier nivel: fonético (sonidos o entonación), morfológico (concordancias, tiempos verbales), sintáctico, léxico-semántico o pragmático-contextual. Los errores de los alumnos se han podido dar en tareas orales dirigidas, redacciones, interacciones de clase... Para estudiarlos se necesitan, obviamente, las producciones de esos alumnos, preferentemente transcritas y anotadas, es decir, convertidas en un corpus, ya sea de tus alumnos de este año, o de los del Cervantes en los últimos 10 años. Por eso, la mayor parte de publicaciones de

análisis de errores lo estudian en corpus (Cestero et ál., 2002; Ferreira et ál., 2014; Pérez Serrano, 2014).

El análisis de interlengua comprende el estudio de las producciones de los alumnos de lengua extranjera, tanto las correctas como las desviadas, el error (Selinker, 1972). Sin embargo, ya que los errores se tratan en profundidad en la investigación basada en análisis de errores que acabamos de explicar, el análisis de interlengua trata aquellas producciones que, sin ser necesariamente erróneas, son comunes en todas las personas que estudian una lengua extranjera, muchas veces, no directamente relacionados con la L1.

Nos referimos a características como la abundancia de pausas, el uso de frases cortas y simples (sin subordinación), uso de términos poco precisos como *cosa*, etc. Para afrontar el análisis de interlengua, igual que en el análisis de errores, se necesitan producciones de alumnos de lengua extranjera, por lo tanto, corpus de los alumnos. A este tipo de corpus le llamamos **corpus de aprendientes** o **corpus de errores**. Y es el único tipo de corpus específico para el aprovechamiento en ELE.

6.2. LOS CORPUS DE APRENDIENTES

Los **corpus de aprendientes** contienen datos de muestras de habla real de los estudiantes de una determinada lengua. En la bibliografía, también se han llamado **corpus de interlengua** o **corpus de errores**. Existen corpus de aprendientes para muchas lenguas meta y desde muchas lenguas de origen de los estudiantes. Por ejemplo, si estuviéramos interesados, podríamos consultar los errores más frecuentes en inglés de los aprendientes hispanohablantes. En el ámbito internacional, se puede encontrar incluso una asociación de estudiosos de corpus de aprendientes que organiza congresos y mantiene recursos útiles en diferentes lenguas (<https://www.learnercorpusassociation.org>).

Los corpus de aprendientes pueden tener, como los corpus en general, diferentes tipos de marcaje y anotación. En primer lugar, suelen estar codificados con la L1 del estudiante y su nivel, y, muchas veces, además incluyen información sociolingüística o del aprendizaje del alumno, como, por ejemplo, si ha pasado un periodo de inmersión en un país hispanohablante y cuánto tiempo.

Además, existen corpus lematizados y corpus en crudo, que son simplemente una recopilación de, por ejemplo, redacciones realizadas por los alumnos de B2 que estudian español en una universidad de Berlín.

Los corpus más anotados contienen los errores marcados en el texto y algunas veces contienen también información sobre el tipo de error cometido. Así, en

una producción como “la gente son” aparecería anotada la palabra *son* como “error morfológico, concordancia”. Y eso posibilitaría realizar búsquedas de todos los errores de concordancia cometidos en un grupo de estudiantes y de esta manera se podría comprobar qué palabras causan más errores de concordancia y en qué nivel dejan de causarlos.

Habíamos dicho ya en un apartado anterior que los docentes e investigadores tienen dos posibilidades: consultar un corpus de errores ya creado, o crear su propio corpus. Pero crear un corpus propio implica recoger datos, digitalizarlos, marcarlos, anotarlos y, después de haber hecho todo eso, empezar a realizar su pequeña investigación. La realización de todo ese trabajo, por ejemplo, en el caso de un trabajo de final de grado, certificaría que el alumno se desenvuelve bien con el método empírico, que sabe realizar todas esas tareas, y, por lo tanto, que es buen investigador, pero ser investigador no es lo mismo que ser profesor. Así que a no ser que se esté preparando un TFM de investigación o una tesis doctoral, recomendamos que se use un corpus ya disponible y solo se cree el corpus de la clase si hay un objetivo concreto que realmente lo haga necesario, como dar correcciones personalizadas a los estudiantes. A esto volveremos más adelante en el libro.

6.3. CORPUS DISPONIBLES

En el caso de que no tengas alumnos de los que recoger datos y quieras hacer una investigación (publicar un artículo, hacer un trabajo de investigación original), hay numerosos corpus disponibles y anotados listos para su explotación.

La parte positiva de usar un corpus ya compilado es que ahorrarás el tiempo que necesitarías invertir en compilar tu propio corpus y anotarlo. La parte negativa es que puede que el corpus no cumpla los criterios que te interesarían.

Por ejemplo, puede que quieras hacer una comparación de los alumnos de dos lenguas y que el corpus no contenga estudiantes de esas lenguas en concreto o que quieras estudiar grupos que hayan aprendido con determinado método y no los incluya o que no diferencie a los alumnos que han estado en inmersión de aquellos que no y esa variable fuera muy importante para ti. También puede ser que quieras estudiar un error en concreto, por ejemplo, errores en el uso del artículo, y que no encuentres un corpus que lo tenga anotado.

Antes de empezar la tarea napoleónica de recoger un corpus para la investigación, siempre merece la pena echar un vistazo a lo que hay en la web. En los siguientes

párrafos vamos a hacer un repaso de los corpus de aprendientes disponibles y sus puntos fuertes. Para cada corpus, hay una referencia citada al final del libro. Esa referencia es siempre un artículo donde se explica el diseño (el proceso de creación), las decisiones metodológicas de creación del corpus y tipología y cómo está conformado, y su tipo de marcaje y anotación. Es decir, al consultar cualquiera de esos artículos, podrás ver la aplicación práctica de todo lo que hemos explicado en el libro hasta ahora y, además, te harás una idea más clara de los puntos fuertes y débiles del corpus, así como de las dificultades a las que se enfrentaron los autores a la hora de crearlos.

La tradición de los corpus de errores, como todo en la lingüística de corpus, comienza con el inglés. En los años noventa del siglo pasado nace el *International Corpus of Learner English* (ICLE) (Granger, 2003) y, a partir de ese momento, empiezan a recoger datos sobre la enseñanza del inglés. El corpus sigue en activo hoy en día; contiene redacciones de alumnos de nivel intermedio y avanzado de 22 lenguas maternas con un total de 5,7 millones de palabras.

Pero pasemos a nuestro interés real. En el caso del español, la tradición empieza con el corpus CEDEL (Lozano, 2009; Lozano & Mendikoetxea, 2013) del cual ya existe una segunda versión en línea. Como la mayoría de corpus, se trata de un corpus de textos escritos, en concreto, de redacciones sobre diferentes temas. Contiene aprendientes con L1 inglés (y un subcorpus más pequeño de L1 griego) y se puede consultar en línea y descargar. Tiene una interfaz cómoda e intuitiva con la que se pueden filtrar las características de los alumnos (figura 5).

Este es un corpus útil para hacer análisis de interlengua, pero no tanto para buscar errores, ya que la interfaz no lo permite. Sí permite, en cambio, buscar por nota concreta de los alumnos o por el título de la redacción, lo que facilita estudios centrados, por ejemplo, en un determinado tipo de texto (descripción, resumen, narración...) o en un tema (persona famosa, viajes...). Este sería un ejemplo de producción de una alumna de A1 para la descripción de una persona famosa:

Denzel Washington es muy guapo. Senor Washington te gusta película. Senor Washington esta alto y mucho tiene deniro. Senor Washington vive en Los Angeles California con el novia en mucho encasa. Denzel Washington encanta football y baseball. Washington Tiene tres ninos. el nino camine Morehouse Universidad y juege football. Denzel Washingtons Pelicuas es John Q, Remember the Titans, y inside man. Me gusta Denzel Washington, el es me favorito actor.

CEDEL2 [Instructions] [Corpus: Spanish natives] [Corpus: L1 English - L2 Spanish] [Corpus: L1 Greek - L2 Spanish]

Search a particular word (e.g., "estación") or a particular filename (e.g., 9_26_0_2_NLP)

Age of learner: 0 - 100 Proficiency level (placement test score): 0 - 100 Proficiency level (self-evaluation): 0 - 6

Age of exposure to Spanish: 0 - 100 Years studying Spanish: 0 - 50 Stay abroad (months): 0 - 300

Other foreign languages: ☐ Japanese ☐ Khmer ☐ Korean ☐ Latin ☐ Malay ☐ Malayalam ☐ Maori ☐ Portuguese ☐ Punjabi ☐ Quechua

Essay title: ☐ 01. ¿Cómo es la región donde vives? ☐ 02. Habla de una persona famosa. ☐ 03. Resume una película que has visto recientemente. ☐ 04. ¿Qué hiciste el año pasado durante las vacaciones? ☐ 05. ¿Cuáles son tus planes para el futuro? ☐ 06. Describe un viaje que has hecho recientemente. ☐ 07. Cuenta una experiencia que hayas vivido. ☐ 08. Habla del problema del terrorismo en el mundo. ☐ 09. ¿Qué opinas de la nueva ley anti-tabaco? ☐ 10. ¿Crees que las parejas gay tienen el derecho de casarse y adoptar niños?

Clear Download Query

Total: 1809 Total Pages: 81

ID	Age of learner	Proficiency level (placement test score)	Proficiency level (self-evaluation)	Age of exposure to Spanish	Years studying Spanish	Stay abroad (months)	Other foreign languages	Essay title
en_wr_9_26_0_2_nlp	26	21	1.75	26	0			02. Habla de una persona famosa.

Figura 5. Interfaz de búsqueda de CEDEL v2

Hoy en día, el corpus que podríamos considerar "oficial" para el español es el *Corpus de aprendices de español* (CAES) realizado por la Universidad de Santiago de Compostela con la ayuda del Cervantes (<http://galvan.usc.es/caes>). Este corpus, como el CEDEL, tiene una interfaz de búsqueda muy lograda para seleccionar las variables de los alumnos. Además, en él se puede buscar por tipo de palabra gramatical al modo de los corpus de lengua general y ofrece comparativas de la frecuencia de la palabra de búsqueda separadas por L1 de los alumnos y por nivel.

Pero en todos estos corpus la búsqueda se realiza por una palabra o colocación, es decir, si bien tienen anotación morfosintáctica, no contienen los errores de los alumnos anotados. Por eso, es especialmente interesante para nuestro cometido el corpus que presentamos a continuación.

En el *Corpus Oral de Español como Lengua Extranjera* (COELE) se encuentran las producciones transcritas de 40 alumnos de español lengua extranjera, una muestra más bien pequeña. Pero el punto fuerte de este corpus es que contiene

anotaciones de los errores, así que permite hacer búsquedas de errores. Estas búsquedas pueden ser por nivel de lengua, por errores de pronunciación, o incluso más específicas, por ejemplo, errores de concordancia de número o errores de tiempo verbal. Los ejercicios de este capítulo permiten explorarlo de manera guiada.

Por supuesto existen muchos más corpus de aprendientes. Algunos de ellos (más bien pocos) son de variedades lingüísticas que no son la española peninsular, como el *Corpus de aprendientes de español como lengua extranjera y segunda lengua* (CAELE/2) de español de Colombia (Hincapié, 2018). También existen versiones especializadas según la L1 de los alumnos como el *Corpus ORal de Aprendientes de Francés* (CORAF), el *Aprescrlor* para aprendientes que hablan neerlandés (Buyse & González Melón, 2012) o el CELEN para hablantes de japonés (Valverde, 2020). Este último no tiene una interfaz web propia donde realizar las consultas, sino que tiene que ser consultado a través de Sketch Engine, que, como dijimos en la sección 2.5, es un programa en el que se puede cargar cualquier corpus (conjunto de textos anotados) para facilitar su consulta y la realización de búsquedas.

En cuanto a los corpus orales, existen muchos menos y, en la mayoría de los casos, no cuentan con audios descargables disponibles para hacer un análisis acústico de los errores de los estudiantes, sino únicamente con las transcripciones (como el de García Gutiérrez (2005)). Eso hace que la mayoría de tesis que se deciden a hacer un trabajo de fin de máster o una tesis de pronunciación de ELE se vean obligados a recoger su propio corpus. Algunos corpus orales disponibles son el *Spanish Learner Language Oral Corpora* de la Universidad de Southampton (Mitchell et ál. 2008) o el C-ORAL-ROM, en versión CD (Llanos et ál., 2007).

Si necesitas un corpus sobre una L1 en concreto, puedes encontrar una lista exhaustiva y actualizada de corpus de aprendientes de diferentes lenguas con su clasificación en <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>. Además, algunos corpus orales de español L2 están disponibles a través de la sección *Second Language Acquisition* de Talkbank, un gran banco de corpus lingüísticos de todo tipo (<https://slabank.talkbank.org/access/>).

El problema es que todas estas bases de datos incluyen solo corpus que cumplen ciertos requisitos de tamaño y anotación: se consideran corpus de verdad los que son ordenados, anotados y reutilizables. Eso hace que para fines específicos existan muchos más corpus que no están indexados. Por ejemplo, Gutiérrez Quintana (2005) creó un corpus de textos escritos por universitarios italianos estudiantes

de ELE, que está en línea y contiene únicamente un PDF de las redacciones con el marcado de nivel, curso y edad, pero sin ningún tipo de anotación o buscador. Como este se pueden encontrar más, de otras L1, en el apartado CORPUS de la web de la revista científica LinRed (<http://www.linred.es/corpus.html>). Se trata, básicamente, de corpus oportunistas como los que podemos crear nosotros a partir de las redacciones de nuestros alumnos.

6.4 PARA MÁS INFORMACIÓN

Núñez Nogueroles, E. (2019). Pasado, presente y futuro de los corpus de aprendientes de ELE. Una revisión bibliográfica. *Reidocrea. Monográfico 2019 – enseñanza de lenguas*, 170-190.

6.5 ACTIVIDADES

6.5.1. Visita la página <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>. ¿Cuántos corpus tienen el español como lengua de estudio (*target*)? ¿Y cómo L1 (*first language*)? ¿Cuál de estas es más interesante para los profesores de ELE? ¿Por qué?

6.5.2. Visita la página <https://slabank.talkbank.org/access/>. ¿Cuál de los corpus orales que hay disponibles te parece más interesante?

6.5.3. Busca en el <http://cartago.llf.uam.es/exist/rest/db/corpus/errores.html> por TIPO DE ERROR “Indicativo/subjuntivo” y LENGUA MATERNA “Francés”. Después de ver esos datos, ¿qué tema repasarías en clase?

6.5.4. Busca ahora resultados de errores de concordancia de número en estudiantes cuya lengua materna es el inglés. ¿Qué errores son más comunes? ¿A qué rasgo lingüístico del inglés se pueden deber (esta pregunta presupone que tienes un nivel de inglés de al menos B1)? Un estudio completo y extenso sobre este tema podría dar lugar a un trabajo de curso: “La concordancia de número en la interlengua en español de aprendientes anglohablantes”.

6.5.5. Ve al CEDEL (<http://cedel2.learnercorpora.com>), y entra o descarga cualquier producción. ¿Qué datos han incluido en el marcaje del corpus?

7

LOS CORPUS EN LA CREACIÓN DE MATERIALES

La preparación de libros de ELE, cuadernos de ejercicios o el material de clase para un curso en concreto (las fichas), sigue un proceso parecido a la creación de cualquier gramática. Así que, como en estas, de manera clásica, nos encontraríamos a un “gramático sabio” encerrado en su despacho decidiendo qué contenidos forman parte de la lengua, cuáles son más básicos y qué ejemplos son perfectos para ellos. Ya hemos visto que, en el mundo real, se ha comprobado que a veces lo que los profesores prevemos como básico y fácil no lo es tanto, que los ejemplos creados por lingüistas a veces tienen poco uso en una situación comunicativa y que cuando más necesitas pensar en léxico de comida o léxico que contenga grafías *g* y *j* es cuando menos se te ocurren. Para hacer todas estas decisiones más fáciles se puede acudir a diccionarios, diccionarios combinatorios, gramáticas, listas de vocabulario y, cómo no, corpus.

7.1. USO DE CORPUS EN LA SELECCIÓN DE CONTENIDO DEL CURSO O LIBRO

El planteamiento de un libro de texto y del contenido o currículo de un curso de lengua (*syllabus*) es muy parecido. En los dos casos, lo primero que hay que decidir son los contenidos que entrarán. Es posible que esta tarea no recaiga nunca sobre nosotros, porque no trabajemos como editores en una editorial de ELE, porque en la escuela en la que trabajamos esa tarea recaiga sobre un coordinador, etc. Pero también es posible que en una escuela pequeña en la que, o bien no había un programa de lengua montado o bien se ha cambiado el claustro, haya que crear el programa. Montar un curso desde cero es una tarea que ocupa muchas horas (aunque se puede aprovechar una parte de un año para otro) y buscar un buen manual siempre ahorra horas de preparación. Pero, por bueno que sea el manual, siempre hay que “hacer nuestro el contenido” y, para lograrlo, lo primero que hay que hacer es tener una panorámica del temario que se va a enseñar en el curso.

Para decidir el contenido que hay que cubrir en un curso o libro, normalmente, miramos el *Plan Curricular del Instituto Cervantes* (PCIC) o, más a grandes rasgos, el *Marco Común Europeo de Referencia para las Lenguas* (MCER o CFER, según si lo citamos por sus siglas en castellano o en inglés) y, a partir de ahí, decidimos. Esa es la manera más rápida y fácil de no equivocarse, aunque hay que tener en cuenta que el PCIC no contiene todo, suele ser más bien una guía general.

Además, no todos los planes de estudios siguen el currículo del Cervantes. Por ejemplo, en las universidades estadounidenses los estudiantes de ELE no persiguen obtener el título DELE o SIELE, sino un grado universitario. Tampoco usan el MCER que, por motivos obvios, no está implantado fuera de Europa. Muchos de los profesores del sistema estadounidense ni siquiera lo conocen y los coordinadores generales no suelen ser proclives a dejar que los profesores de su sistema usen un currículo europeo para decidir su *syllabus*. Sin embargo, muchas universidades chinas sí se guían por el MCER porque tienen en mente el DELE.

En casos como esos, la decisión o la justificación de qué incluir en un programa y en qué nivel hacerlo se puede hacer a partir de corpus (Caballero et ál., 2012). A través de ellos, se puede ver qué situaciones comunicativas y recursos lingüísticos son más frecuentes y, por lo tanto, interesantes para incluir en un método o en un curso de ELE.

La idea de decidir qué incluir en un programa o justificarlo mediante corpus es buena, pero ¿cómo se puede buscar el contenido que debemos incluir en un determinado nivel en un corpus? La respuesta no es fácil. Para empezar, el contenido que queremos buscar puede ser: a) **gramatical** (¿debería incluir el imperativo en un A2?), b) **léxico** (¿incluyo vocabulario sobre las vacaciones en A1?) o c) **funcional** (¿en qué nivel incluiría ofrecer una justificación?). Los corpus, tal y como los hemos explorado hasta ahora, nos pueden servir para los dos primeros tipos de contenido, pero no para el tercero. En cambio, hoy en día, la mayoría de los programas de ELE se estructuran alrededor de funciones comunicativas.

Hasta donde llega nuestro conocimiento, solo existe un corpus en el que se puede buscar por función comunicativa, el *C-Or-Dial* (Nicolás Martínez, 2012, 2015), y solo se puede realizar esa búsqueda en su versión en CD-ROM y no en su web. En la práctica, que no esté disponible en línea significa que no es posible elegir qué funciones comunicativas incluiremos en un curso directamente desde un corpus (a no ser que nos leamos el corpus completo para hacernos una idea

de las más frecuentes). Aunque, como decíamos al principio, esto no supone un gran problema porque siempre tendremos el PCIC para saber qué funciones incluir.

En cambio, una vez que tenemos algunos límites de contenido establecidos, por ejemplo, el registro que queremos tocar, podemos buscar un corpus especializado en el que aparezca ese tipo de lengua. Por ejemplo, en un tema de dar la opinión, podemos usar un corpus de opiniones de Amazon y un corpus de cartas formales nos puede servir para extraer los saludos más frecuentes, pero esto lo veremos con más detalle en el epígrafe siguiente.

Volvamos a la selección de contenido de un curso. Hay un tipo de cursos para el que se tiene una idea clara del campo que se quiere tratar y muchas veces no se cuenta con un currículo claro: el español para fines específicos. El **español para fines específicos** tiene varias particularidades que hacen que no se pueda aplicar un plan general como el PCIC. Estas peculiaridades se encuentran, básicamente, en la terminología y el registro que se enseña. Vamos a ver por qué.

La **terminología** (vocabulario) que en un curso general no se vería hasta niveles altos o incluso no se vería en absoluto puede tenerse que dar en un curso de fines específicos en niveles bajos. Por ejemplo, para un médico, una palabra como *vía* (en la colocación *vía intravenosa*) o *bisturí* formarán parte de su vocabulario más usado y, sin embargo, no entran en ningún programa de ELE. Otro ejemplo, un curso de español a una cadena de panaderías que quiere entrar en el mercado español, a sus empleados no les bastará con conocer la palabra *harina*, sino que necesitarán, además, sus tipos (*de fuerza, de repostería, de centeno...*). Además, las diferentes profesiones usan jergas, a veces, por necesidad, como en los ejemplos de léxico que acabamos de ver, pero otras, simplemente por convención, por ejemplo, los médicos usan *molestias* en vez de *dolor* y a primera vista no hay una explicación del por qué (Cordella, 2002).

En los cursos de lengua general, se suele enseñar un **registro** estándar, con algunas tendencias al formal (por ejemplo, cuando se trabajan cartas de presentación) o al coloquial (cuando se hacen intervenciones en clase). Pero en uno de español para fines específicos, por ejemplo, en un curso de español de los negocios, a los alumnos solo les interesará el registro más formal; en cambio, en uno de español para el turismo, se buscará un lenguaje más cercano.

Todo esto hace que sea oportuno usar corpus, en concreto, un corpus de lengua especializada.

Para poder establecer nuestro propio plan de estudios para fines específicos podemos acudir a corpus especializados, pero la mayoría de ellos no están disponibles con un buscador en línea como los que hemos visto hasta ahora. Por ello, resulta mucho más cómodo acudir a los estudios que se han realizado a partir de esos corpus. En ellos, se pueden encontrar listas de léxico y reflexiones sobre su discurso ya preparadas.

Por ejemplo, a partir del análisis del *Corpus Multilingüe de Turismo de la Universitat de València* (COMETVAL) se ha podido confeccionar un diccionario y publicar numerosos estudios. En estos últimos se pueden encontrar patrones de denominaciones como *turismo* + adj o *turismo de* + nombre: *turismo de aventura*, *turismo activo*, *turismo rural*. Y también léxico de uso habitual en español del turismo, como *quad*, *avistaje de aves*, *hidrotrineo* o *escafandrismo* (López Santiago & Giménez Folqués, 2017).

Si no existe un corpus del tipo concreto de lengua que queremos enseñar (que se sepa, no existe un corpus de lengua de los panaderos o de los albañiles, por poner algún ejemplo), se pueden usar manuales o glosarios. Pero enseñar el léxico de una profesión suele ser insuficiente, las **jergas** suelen incluir también otros rasgos. Por ejemplo, en derecho eliden la conjunción *que* en subordinadas de subjuntivo (dicen *solicito sea tenido en cuenta* donde deberían decir *solicito que sea tenido en cuenta*), algo que no es normativo en la lengua general, también hacen un uso más abundante del subjuntivo, de las subordinadas y de las formas impersonales que en la lengua general. Además, mantienen formas arcaicas como *fallo* por *yo hallo* o el futuro de subjuntivo *si no comparecieren*. Así que también deberemos tener en cuenta la sintaxis cuando enseñemos lenguas especializadas.

Otra posibilidad es crear nuestro propio corpus a partir de muestras de la lengua especializada que queremos enseñar. Para hacer esto, Vangehuchten (2004) muestra un método bastante rápido (si se tienen conocimientos de informática/ programación suficientes) y eficaz. Él propone seleccionar los manuales académicos de introducción a la disciplina (por ejemplo, *Manual de las ciencias del deporte* o *Introducción al derecho romano*) más usados en el ámbito y, a partir de esos manuales, seleccionar el vocabulario y las formas propias de la materia. En su artículo, Vangehuchten (2004) presenta una manera matemática de seleccionar ese vocabulario y nos recuerda que, según la bibliografía, el conocimiento del 95 % del léxico de un texto asegura su correcta comprensión y la capacidad de extraer significados desconocidos a partir del contexto, mientras que un conocimiento del 98 % del léxico proporciona una lectura cómoda.

En este epígrafe nos hemos centrado en aquellos casos en los que hay que montar unidades desde cero, pero estos no son los únicos en los que podemos tener que decidir qué contenido incluir. Cada clase (cada unidad didáctica ya creada) necesita adaptarse al contexto de nuestro grupo concreto y eso también significa elegir qué formas lingüísticas se van a ver.

7.2. QUÉ FORMAS LINGÜÍSTICAS INCLUIR EN CADA UNIDAD DIDÁCTICA O EN CLASE

Las situaciones comunicativas que se incluyen en una unidad condicionan las formas lingüísticas que se enseñarán. Sin embargo, siempre hay un espacio para la elección. Casi todos los métodos de ELE de A1 empiezan por una unidad 1, o incluso 0, en la que se ven los saludos y las presentaciones. En esa unidad, normalmente se incluyen formas como *hola*, *buenos días* / *buen día* o *buenas tardes*. Sin embargo, en casi todo el mundo hispanohablante, hay una forma que los alumnos oirán mucho pese a ser un préstamo: *ciao*, *chao* o *chau*. Esta forma, probablemente, no es la primera en la que uno piensa al montar una unidad, pero sí es una forma actual y muy frecuente.

Cuando uno se enfrenta a crear una unidad, después de haber elegido la situación comunicativa o la función que va a incluir, el paso siguiente consiste en buscar las formas lingüísticas que se incluirán. Para ello, un acercamiento natural sería el siguiente. Pongamos que queremos que el alumno sea capaz de describir la imagen de una persona. Una vez tengamos decidido eso, nuestra intuición de docentes y lingüistas (y no nuestra intuición como nativos) nos indica que, para hacerlo, el alumno necesita el presente de indicativo y léxico de descripción física. De ahí, escogeremos las formas que nos parezcan más adecuadas al nivel (o de nuevo acudiremos al PCIC). Ahora, para comprobar las ventajas, los problemas y dificultades que conlleva, vamos a ver el mismo proceso con un ejemplo en el que se ha confiado en corpus en vez de en nuestra intuición para realizar la misma tarea.

Imaginemos una unidad que tiene como objetivo final “ser capaz de desenvolverse en una estación de tren”. Caballero et ál., (2014) proponen un acercamiento de corpus para llegar a cumplir esa tarea. Para definir las formas lingüísticas que se incluirán en la unidad, Caballero et ál. (2014) ofrecen un análisis de las expresiones en las taquillas usadas por los nativos, InfoTren, y una comparación con un corpus de simulaciones (*role-play*) tanto con nativos como no nativos, FerroviELE. Ese análisis les permite descubrir que, contrariamente a lo que se pudiera pensar, los

no nativos usan un léxico más variado y mayor vocabulario de viajes. Y también un mayor nivel de cortesía y menor número de interacciones (debido a que se interrumpen menos). Además, el análisis del mismo corpus muestra que los nativos usan pausas llenas, sobre todo, *eh* y *ajá*, mientras que los extranjeros usan *um*, *uh* y *em*.

Pero vamos a ver en (2) un ejemplo de InfoTren de una intervención real, para que analicemos el tipo de respuesta a la que se enfrentaría un estudiante que llega a Valencia y pregunta en español el precio de los billetes de tren a Castellón.

- (2) Permita un momentito que le confirme el precio de Castellón, a ver, que me parece que no se lo he dado bien.

Supongo que estaremos de acuerdo en afirmar que (2) es un ejemplo de lengua oral real y bastante prototípica, pero ¿se parece a lo que se enseña en nuestras aulas? Parece claro que no, sin embargo, ese es el escenario que se van a encontrar los estudiantes cuando lleguen a un país en inmersión. Las situaciones perfectas que han trabajado en clase les servirán para comunicar, pero no para entender, porque las respuestas de sus interlocutores serán de lo más variopintas e inesperadas.

En general, los corpus contienen expresiones naturales y cooperativas (para eso son muestras reales de habla y no inventadas o creadas *ad hoc*), aunque son muestras difíciles de entender para un extranjero. Pero, ¿se puede llevar este tipo de lengua al aula? En realidad, sí. Está claro que no vamos a incluir una frase tan compleja como la de (2), pero muchas de las estrategias que usa la vendedora sí pueden ser temario “llevable” a una clase de ELE.

De esta producción, podemos extraer cosas tan usadas y poco incluidas en los manuales de español como: el uso del diminutivo como atenuador y cortés en una situación formal (*un momentito*), expresiones de alta frecuencia como *a ver* o el marcador *que* en *que me parece que no se lo he dado bien*. De ellas, probablemente la más fácil de incluir y útil sea el diminutivo. Vamos a verlo con más calma. En el contexto de una cafetería, cuando se pide el café de una manera especial, muy frecuentemente se hace con el diminutivo, *café con la leche calentita*, *templadita* o *con un hielito*. El emisor sabe que está pidiendo algo de más y compensa atenuando la petición mediante el diminutivo y ello no hace el café o el hielo más pequeño.

Por lo tanto, una búsqueda rápida en cada unidad puede aportar recursos que no aparecían en el manual, pero que son relevantes. Por ejemplo, si damos clase en España, ver cómo piden los nativos un café o unas patatas bravas puede ser de especial interés. ¿Cómo y dónde buscarlo? Si hacemos la búsqueda *bravas* en el

CORPES (recuerda, en este caso, como forma), muy pronto encontrarás líneas de concordancias como: *¿Y si nos pone dos de bravas? Y algo de pan también, por favor*. De este caso, es interesante la estructura *poner* + número de algo como una estrategia útil y muy común que se puede incluir. Por otro lado, descartaríamos la segunda frase, porque el adverbio *también* tiene posición no prototípica que obliga a un dominio muy alto de la entonación, ya que para producirla correctamente hay que producir *pan-también* en el mismo grupo entonativo (sin pausa entre las dos palabras).

Otro ejemplo, pongamos que queremos incluir más formas de invitar. Además de al CORPES, se puede acudir a un corpus oral como Valesco o el COLA, que son corpus coloquiales. Y a partir de ahí, tenemos que pensar en una palabra clave. Si queremos saber como se invita en España a tomar el café, podemos buscar *café* en Valesco y dos ocurrencias serán invitaciones: *¿Nos vamos a tomar café?* / *¿Un café o qué?*. Hasta aquí las ocurrencias, lo que incluyas en el manual o la clase, en qué nivel lo hagas y cómo modificarlo para eliminar aquellos rasgos lingüísticos que no se están tratando en ese momento del curso, queda a nuestro criterio como profesores.

7.3 USO DE CORPUS EN LA CREACIÓN DE EJEMPLOS

La creación de ejemplos es una de las tareas más creativas en la elaboración de materiales de ELE. No pecar del abuso de expresiones del tipo *Juan come manzanas* a la vez que se mantiene una estructura sencilla con léxico adecuado al nivel (y muchas veces al contenido del tema) y que sea de uso común no es tan fácil como podría parecer a simple vista.

Lo mismo sucede en el ámbito de la clase: cuando más necesitamos un ejemplo, la mente se queda en blanco y acabamos diciendo una barbaridad o soltando un tópico. Y también, otras veces, al preparar una clase, dudamos de si nuestros ejemplos son adecuados. Si no estamos seguros de si algo se usa en el español general, es de nuestra variedad o incluso de si algo es correcto, podemos acudir a un corpus para comprobar si eso “se dice”. Esta idea volverá a aparecer en el capítulo 9 cuando hablemos de los corpus para la corrección, específicamente, de cómo corregimos las producciones de nuestros alumnos.

Para ver cómo podemos crear ejemplos a partir de corpus, volvamos a *Juan come manzanas*. Una frase de este tipo ha podido surgir porque necesitamos una estructura sujeto + verbo transitivo + objeto directo, así que, ¿por qué no buscar eso en un corpus y dar a los alumnos una forma que tenga muchas ocurrencias?

Si haces la prueba, verás que algo que sí decimos mucho en español es *X no come carne* o *X come alimentos* + adjetivo, algo como *poco sanos* o *con mucha grasa*. Un ejemplo de esto sería *Su bebé come alimentos sólidos*. Puedes comprobarlo con una búsqueda en el CdE buscando el lema *comer_V** con cualquier nombre como colocado (usando [n*]).

Los corpus nos proveen de numerosos ejemplos que se han usado en la vida real en algún momento. La contrapartida de esto es que es posible que muchos de los ejemplos que obtengamos no sean adecuados al nivel. Por suerte, tendremos tantos ejemplos que podremos pasar directamente a los que sí sean adecuados y, además, los ejemplos del mundo real suelen ser más divertidos que los que se obtienen en libros de texto.

Como siempre, tenemos que conocer a nuestra clase para saber qué tipo de ejemplos podemos llevar. Pero en los libros de texto, muchas veces la corrección política o la necesidad de adaptarse a todo tipo de públicos, por ejemplo, evitando que aparezca alcohol en los libros, hace que no podamos usar como ejemplos *Se iba de cañas los viernes después del trabajo* o el par mínimo² *pero-pedo*.

Por todo ello, cuando hablamos de usar corpus en la creación de ejemplos, la idea no es tomar directamente una línea de concordancia que nos dé el corpus, sino que muchas veces es más recomendable usar el ejemplo obtenido del corpus como inspiración y modificarlo para que incluya el léxico o formas gramaticales que necesitamos.

7.4. PARA MÁS INFORMACIÓN

Caballero, M.; Díaz, L. & Taulé, M. (2012). Cómo prever y organizar los contenidos lingüísticos y sociopragmáticos de un escenario comunicativo en clase de E/LE a partir del uso de corpus. *Monografías MarcoELE: Estudios sobre aprendizaje y adquisición de lenguas y sus contextos*, n° 15.

7.5. ACTIVIDADES

7.5.1. Estás preparando una unidad didáctica de A1 que tiene como tema “la familia” y has llegado al límite de piezas léxicas recomendadas para una sesión. Tienes que elegir entre las palabras *primo* y *sobrino*. ¿Cuál elegirías? Justifícalo por su frecuencia en corpus. Tendrás que

2 Un **par mínimo** es un par de palabras que se diferencian por un único sonido. Se suelen usar en los ejercicios de pronunciación para ayudar a los estudiantes a distinguir entre dos sonidos como *r-rr* con *coro-corro* o *p-b* con *pala-bala*.

elegir un corpus lo suficientemente grande, para que tus resultados sean representativos.

7.5.2. Busca en el CdE (NOW) la palabra *falda* y obtén la lista de concordancias.

1. Observa las 20 primeras concordancias. ¿Cuáles de los adjetivos y colocaciones incluirías en un libro de nivel A1 y en uno de nivel B2? Nos referimos a expresiones del tipo *falda de lunares*, *falda larga*, etc. En algunos casos, las colocaciones no se podrán obtener, puesto que la palabra *falda* aparece sola.
2. Repite la búsqueda esta vez para obtener los colocados más habituales de *falda* + adjetivos: en la parte derecha (configura colocados a la izquierda 0, a la derecha 1). ¿Cuántos de los 20 primeros incluirías en un A1?
3. ¿Y si estuvieras preparando un curso para los empleados de una cadena de moda que quiere vender sus diseños en España?

7.5.3. Ve al corpus Valesco (<http://www.valesco.es/corpus/corpuusrch.php>) y realiza una búsqueda por intervención con el contenido *es que*. ¿Cuál es la función de las ocurrencias de *es que* en el corpus?

7.5.4. Tus alumnos te han pedido una clase monográfica de comida. Elige un ingrediente o un plato típico (por ejemplo, *tortilla* o *huevo*) y extrae de un corpus sus colocaciones/coapariciones para obtener el contenido léxico de la clase. Si usas el CORPES, no olvides ordenar por frecuencia para poder ver las coapariciones más habituales. Tienes dos enfoques posibles para la clase: 1) expresiones fraseológicas (*darle la vuelta a la tortilla*, *costar un huevo*); 2) colocaciones (*tortilla de maíz*, *huevo pochado*).

7.5.5. Busca en el CdE la palabra *cuando* seguida de verbos en imperfecto. Elige una (la que te parezca más útil para llevar a clase) de las tres colocaciones resultantes de más frecuencia y mira las 10 primeras líneas de concordancias. Adapta tres líneas de concordancias para que los ejemplos obtenidos sean aptos para llevar a clase.

8

ACTIVIDADES CON CORPUS PARA LLEVAR AL AULA

8. ACTIVIDADES CON CORPUS PARA LLEVAR AL AULA

Hasta ahora hemos visto cómo los profesores pueden usar corpus para investigar sobre el ELE y para crear materiales y ejemplos, pero no hemos visto actividades en las que los alumnos sean los que tienen que usar los corpus. Eso es lo que haremos en este capítulo.

Los corpus tienen, en general, mala fama dentro del mundo del ELE. Se los considera instrumentos de investigación, solo útiles para aquellos que quieran realizar una tesis doctoral sobre el español lengua extranjera en determinado grupo de alumnos. De hecho, en las publicaciones en las que se habla de corpus y ELE, el contenido se suele centrar en esas tareas y son muy pocos los estudios disponibles donde realmente se hace una aplicación de los corpus al aula. Incluso cuando se hace, se tiene la preconcepción de que solo se pueden usar en niveles altos. Sin embargo, el uso de corpus en clase puede ser útil e incluso divertido y, para demostrarlo, en los siguientes epígrafes vamos a ver una serie de actividades que se pueden realizar en clase con corpus.

Las actividades están clasificadas según el tema que tratan o la metodología que usan, pero este no sería el orden en el que lo presentaríamos en una clase o en un curso. El orden de presentación de las actividades (secuencia didáctica) y el proceso de introducir corpus en un curso se tratará en el capítulo 11.

8.1. ACTIVIDADES SOBRE CULTURA

La lengua, específicamente el cambio lingüístico (la manera en la que evolucionan las lenguas), es una muestra preciosa de los cambios que sufre la sociedad. Es por ello por lo que, si consultamos en un corpus con datos históricos la palabra *ordenador*, podremos deducir cuándo se amplió su significado de algo que ordena cosas a ese aparato electrónico indispensable para todos nosotros.

Esto aplicado a la cultura de un país nos puede servir para introducir temas sobre diferencias culturales o cambios sociales. Una búsqueda en un corpus de noticias

sobre la colocación *matrimonio homosexual*, nos puede servir para hablar en clase del estado judicial del matrimonio igualitario en diferentes países hispanohablantes, o una búsqueda de la palabra *aborto* nos puede dar pistas de cuándo ha habido un debate real sobre el aborto en diferentes países.

En las actividades del capítulo 3, usamos Google Ngram para ver la aparición y el aumento en el uso de la palabra *médica* (como sustantivo, lo que excluye los casos de *revisión médica* y similares). Esa misma actividad es fácilmente adaptable a una clase, como primer contacto con actividades de corpus. El profesor puede enseñar los resultados en clase abierta, y eso puede dar pie a un debate sobre la aparición de las primeras licenciadas en medicina (1857 en España) y a una comparación con cuándo se pudieron licenciar en medicina las mujeres en sus países de origen. Otra opción es presentar el debate existente en español sobre si las mujeres licenciadas en medicina prefieren que las llamen *médico* o *médica*. De manera más global, se puede aplicar al debate existente con todas las palabras con femenino de nueva creación como *presidenta* o *arquitecta*, o incluso la polémica sobre *portavoza* con menciones a las recomendaciones de las Academias y el uso real en la calle.

Es importante recordar las características de cada corpus para saber dónde efectuar las búsquedas. Por ejemplo, si queremos buscar ocurrencias de *portavoza*, tendremos que acudir a un corpus monitor, como la sección NOW del CdE, y preferiblemente que incluya muchos textos periodísticos, puesto que el debate sobre el uso del término se dio en los medios de comunicación. En cambio, si busco *portavoza* en Google Books, no lo encontraré porque la palabra no ha aparecido en ningún libro publicado hasta 2012.

Es importante recordar también que, pese a todo, estamos dando una clase de ELE y no una clase de historia de la lengua o una de corpus, por lo tanto, algunas de las reflexiones que hemos hecho en este curso, por ejemplo, preguntarnos a qué se debía la aparición de *médica* en la Edad Media o de *fuerte miente* en la actualidad no tiene, en principio, una aplicación directa a la clase.

Vamos a ver un ejemplo de ejercicio posible con Google Ngram. Pongamos que estamos trabajando una unidad en la que, como parte de la cultura española o de los motivos para estudiar español, se incluyen el flamenco y los toros. Y quiero poner en contexto estas dos aficiones. Puedo pedir a mis alumnos que busquen en Google Ngram palabras como *flamenco*, *cantaor*, *toros*, *torero* y *corrida*. A partir de su búsqueda, pueden pensar, en parejas, qué aficiones han ganado adeptos y perdido adeptos en la historia. El resultado de la búsqueda conjunta

de los dos términos es el que aparece en la figura 6. En ella se puede ver cómo las publicaciones en libros en las que aparece el término *flamenco* han ido aumentando desde los años ochenta. A la vez, las apariciones de los *toros* en las novelas y ensayos han disminuido. Estas tendencias se pueden discutir a la luz de los cambios en la sociedad y movimientos ciudadanos como el animalismo que seguramente también se pueden encontrar en los países de origen de nuestros estudiantes. Como curiosidad, el pico de la palabra *toros* (también se observa en *torero*) que se ve entre los años 1875 y 1883 coincide con la época del estreno de la ópera *Carmen* de Bizet (1875), una época en la que las corridas de toros eran especialmente populares.

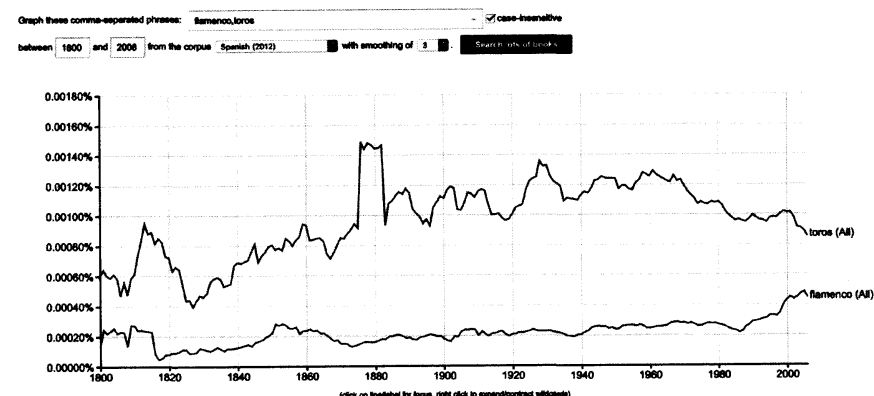


Figura 6. Ocurrencias de flamenco y toros en Google Ngram Viewer hasta el año 2006

A continuación, hay una propuesta de enunciado para el ejercicio de arriba. En él, se incluye 1) una pequeña contextualización de la herramienta, que en clase se podría ampliar con una búsqueda de ejemplo en clase abierta, y 2) las instrucciones de búsqueda en el corpus, que en este caso son muy sencillas.

Google Ngram es una web que nos permite ver cuánto aparece una palabra en los libros publicados por años. Ve a <https://books.google.com/ngrams/> y busca las palabras *toros* y *flamenco*.

- ¿Qué palabra se usa menos que antes? ¿*Flamenco* o *toros*?
- ¿Por qué crees que es?
- Comenta con tu compañero: ¿en vuestros países pasa lo mismo?

8.2. ACTIVIDADES DESDE LA GRAMÁTICA-TRADUCCIÓN

La **gramática-traducción** es, probablemente, el método de enseñanza de lenguas más antiguo del mundo (Melero, 2000). Consiste en estudiar la gramática de la lengua para así poder realizar traducciones de obras clásicas y es el método de aprendizaje de lenguas que se sigue usando en la actualidad para la enseñanza de lenguas muertas como el latín y el griego clásico. El hecho de que el objetivo de este tipo de aprendizaje sea ser capaz de realizar traducciones y no de hablar la lengua hace que este método no goce de mucha popularidad. Pero, aunque los ejercicios de gramática-traducción no gocen de mucha salud en la actualidad, pueden interesar a los estudiantes en casos concretos, por ejemplo, para trabajar obras literarias. Estos pueden ser una buena salida como ejercicio de traducción en clases especializadas o universitarias, por ejemplo, de traducción de poesía o en casos de clases privadas en las que el alumno no tiene un día especialmente comunicativo.

Existen numerosos materiales para trabajar ELE con textos literarios, pero normalmente se plantean unidades didácticas en las que la comprensión del texto y el trabajo del vocabulario se hace enteramente en la lengua extranjera. La elección de los textos que se incluyen en estas unidades es de las tareas más difíciles que se pueden abordar en la creación de materiales; elegir un texto rico y a la vez adecuado al nivel no es nada fácil.

Sin embargo, hay un tipo de textos que no son para nada adecuados al nivel, pero que nos puede interesar incluir en clase. Me refiero a textos de especial valor literario que, en ocasiones, son incluso el motor que motiva a nuestros estudiantes a estudiar español. Se dice, por ejemplo, que personas de renombre, como Pushkin o Freud, aprendieron español para leer el *Quijote*, aunque también puede ser, por supuesto, que fueran unos *posers*. Además, muchas veces nos encontramos que durante el curso se celebra un centenario de conmemoración de cierto autor o un ciclo especial, por ejemplo de poesía, dentro de la escuela o universidad, que nos obliga a introducir obras clásicas en el temario.

No obstante, llevar a clase obras de la complejidad lingüística y estilística del *Quijote* o *Cien años de soledad* puede ser una locura. Sin contar que, sin la suficiente explicación sobre la lengua, nuestros estudiantes pueden acabar hablando de irse de vacaciones a una *ínsula* como Sancho Panza, en vez de a una *isla* como en español del siglo XXI, o tratándonos de vos.

Para paliar esto, se pueden ver las obras junto con la traducción. También de manera cultural, puede ser interesante consultar las diferentes traducciones de obras extranjeras al español, para que los alumnos vean diferencias de traducción según la variedad lingüística o la época en que se realizó. Este tipo de actividades pueden ser interesantes sobre todo en clases universitarias en las que se suele estudiar la lengua más ligada a la literatura.

Los ejercicios de traducción y comparación de la obra original con la traducción pueden ser una buena idea, especialmente en el caso de textos míticos en español, pero que no se pueden llevar a clase sin adaptar a causa de su dificultad. Obras como las mencionadas arriba entrarían dentro de esta categoría y se pueden tratar en clase con fragmentos muy controlados como, por ejemplo, los inicios que, justamente en el caso de estas dos novelas, son especialmente famosos. Otra opción viable para clase sería trabajar con microrrelatos, como el famoso “Cuando despertó, el dinosaurio todavía estaba ahí” del escritor hondureño Augusto Monterroso, que presenta un ejemplo perfecto del uso del par imperfecto-indefinido para que los alumnos vean cómo se traduce a su lengua esta diferencia de aspecto verbal.

Para poder hacer esto, necesitamos un corpus paralelo donde aparezca la obra que vamos a ver en clase. Para buscarlo, se puede consultar la base de datos OPUS (<http://opus.nlpl.eu>), que contiene una lista de corpus paralelos donde hay corpus de textos literarios, entre muchos otros. Pero es muy posible que no encontremos la obra que buscamos en OPUS o que nos interese trabajar con una edición en concreto porque es la que han comprado nuestros alumnos. En ese caso, siempre podemos crear nuestro propio corpus o... mandar a nuestros alumnos que lo hagan ellos. Vamos a ver un ejemplo de actividad usando justamente este método.

Pongamos que damos clase en un curso universitario, cuyo programa es un “Reading Texts”, esos cursos en los que se trabaja la lengua a partir de textos literarios. Queremos trabajar al mismo tiempo el microcuento hispanoamericano y los pasados. Para hacerlo, decido usar el microcuento de Monterroso que hemos mencionado antes, *El dinosaurio*. Puedo organizar una clase invertida en que los alumnos, en casa, tengan que recopilar traducciones del microcuento (de ediciones actuales, como la de Texas University Press, pero también resultados que encuentren en internet, que hay muchísimos). En clase, en grupos, tendrían que crear una traducción propia. Y, cuando la tuvieran lista, en clase abierta, se podría reflexionar sobre la elección de tiempos verbales o de adverbios que han hecho y por qué. Este ejercicio tiene especial jugo en inglés, puesto que la traducción literal “When he awoke, the dinosaur was still there” convive con otras que intentan

captar aún más el aspecto imperfectivo de la segunda parte. Esto lo hacen eligiendo verbos o perífrasis cuyo aspecto léxico (porque el inglés no expresa el aspecto en la desinencia del pasado) es especialmente imperfectivo, es decir, que enfatizan el hecho de que la primera acción está acabada (la de despertar) y la segunda no (la de estar en un sitio). Para hacer esto mismo, en español, usamos perífrasis de gerundio (*estar* + gerundio, *andar* + gerundio, *seguir* + gerundio, *continuar* + gerundio). Otras traducciones, en cambio, traducen: "On waking, the dinosaur was still there", donde la primera parte pierde el valor perfectivo del español.

Este tipo de material puede ser muy rentable si das clase en países no hispanohablantes y no solo universitarias. Como hemos dicho, muchos estudiantes se ven atraídos por la literatura escrita en español y una actividad como esta puede servir para hacer algo diferente.

8.3. ACTIVIDADES CON ENFOQUE COMUNICATIVO

El **enfoque comunicativo** (Canale & Swain, 1980) estableció como objetivo principal de la enseñanza la comunicación efectiva en la L2, priorizando las funciones de la lengua (funciones comunicativas y pragmática) por encima de la estructura (gramática). La comunicación oral se trabaja, en este enfoque, a través del uso de actividades en grupo, juegos de rol, entrevistas entre estudiantes y "rellenahuecos" (vacíos de información), mientras que la gramática se enseña a partir de su función comunicativa, siguiendo un **aprendizaje inductivo** y fomentando el **aprendizaje por descubrimiento** (puedes buscar estos conceptos en el **Glosario**, página 147).

Como se puede observar, gran parte del trabajo de este método se centra en hacer hablar a los estudiantes, algo para lo que los corpus no nos serán de gran utilidad. No obstante, los corpus nos pueden ayudar a mejorar las actividades enriqueciéndolas con contextos reales y a explicar gramática de manera inductiva a partir de los propios descubrimientos de los alumnos. Vamos a ver ejemplos de estos dos usos.

Para introducir los corpus en las actividades, nos podemos remitir a un trabajo clásico del método. En él, Blum-Kulka (1980) proponía contextos a sus alumnos para que respondieran con una estructura que se pudiera usar en esa situación. Al modo de estas actividades, se pueden proponer situaciones comunicativas a los alumnos para que respondan y después comprobar en un corpus cómo han respondido los nativos. Para ello, se puede usar un corpus de situaciones como el mencionado FerroviELE, pero también corpus orales como Val.Es.Co, que contienen situaciones divertidas que se pueden comentar en clase.

Como decíamos, el método inductivo de aprendizaje es el gran eje vertebrador de la enseñanza de gramática en el método comunicativo. Y esto hace que los corpus sean fácilmente introducibles en un aula. Para verlo (esquema 6), vamos a pensar en una secuencia típica de clase para explicar el indefinido:

El alumno ve un video o lee un texto donde aparece la gramática que se va a usar, por ejemplo, una biografía.

El alumno subraya todos los tiempos verbales y reflexiona sobre a qué tiempo se refieren. Posiblemente, rellena una frase sobre el uso del tiempo verbal.

El alumno rellena un cuadro en el que faltan dos o tres formas del indefinido que puede encontrar en el texto (quizá las de tercera persona).

Empieza una secuencia de ejercicios desde los vacíos de información (más formales) hasta actividades más abiertas como escribir la propia biografía o contársela al compañero.

Esquema 6. Modelo de secuencia didáctica que aborda la gramática de manera inductiva

¿Dónde encajarían los corpus aquí? Los corpus son el *input*, sustituyen al texto del manual o preparado por el profesor. En el lugar de la parte 1, los alumnos pueden buscar casos en corpus para después, en pareja o grupo, inferir las reglas gramaticales. Para hacerlo sirve cualquier corpus que devuelva concordancias como el CORPES o el *Corpus del español*. El nivel donde se puede aplicar depende totalmente de la tarea que se prepare.

En las actividades ya se han abordado algunas de ellas. Para el caso de extracción de reglas gramaticales, algunos ejemplos pueden ser obtener la norma de alternancia de *y/e* y *o/u* a partir de una búsqueda de las dos conjunciones. Esta es especialmente interesante porque se puede hacer en niveles muy bajos y puede servir para que los alumnos aprendan a buscar corpus. Además, es una búsqueda tan fácil que se puede realizar incluso en un buscador general como Google, usando las opciones de búsqueda literal. Para hacerlo, buscamos las conjunciones entrecomilladas y con espacios delante y detrás de la palabra): para una actividad se busca, por un lado, "y" y, por el otro, "e"; y, para otra actividad, "o" y "u".

Otro fenómeno interesante que se puede explicar con corpus es la alternancia entre indicativo y subjuntivo. En este caso, la manera más controlada de que los alumnos puedan extraer una norma clara a partir de datos es realizar búsquedas de *creo que* y *no creo que*. La actividad consistiría en que los alumnos apunten qué modo aparece detrás de cada una y expliciten la regla gramatical que subyace. Otros fenómenos que se pueden incluir de esta manera son la alternancia de *ser* y *estar*, *por* y *para* o el uso de los adverbios negativos, que aparece en las actividades de este tema.

En el caso del léxico, aunque trataremos detenidamente todas las posibilidades en el epígrafe siguiente, la actividad más común sería que, a partir de las líneas de concordancia, los estudiantes tengan que deducir el significado de la palabra, por ejemplo, buscar *pollo* y deducir que se trata de algo que se come. También se pueden plantear preguntas más concretas, por ejemplo, a partir de la búsqueda de la palabra *maravilloso*, los estudiantes podrían deducir si tiene un significado positivo o negativo. O para la palabra *tenedor* dilucidar si se trata de algo que se come o un utensilio para comer.

Las actividades de corpus responden muy bien a la esencia de la filosofía del método comunicativo porque en ellas se destaca el papel activo del estudiante, que es quien debe a) efectuar la búsqueda, y b) descubrir la regla que subyace. Además, ese aprendizaje por descubrimiento es a partir de textos reales de español que han producido nativos, por lo tanto, no es una figura autoritaria (el profesor) quien dicta una norma, sino que el estudiante ve que es así como la gente lo dice.

8.4. ACTIVIDADES DE ENFOQUE LÉXICO

El enfoque léxico normalmente no se considera una metodología diferente del método comunicativo, sino una tendencia en la que se da más importancia al léxico. Por lo tanto, los principios que subyacen son los mismos que hemos visto en el epígrafe anterior con el añadido de que en este enfoque se da especial relevancia a tareas como la **derivación** (la creación de palabras a partir de morfemas derivativos), a aprender las palabras en constelaciones semánticas (por **familias léxicas**) y, en especial, al aprendizaje del léxico en algo que en las primeras propuestas llamaban *chunks* (Lewis, 1993). Estos *chunks* serían en la terminología que hemos manejado hasta ahora nuestras llamadas **unidades fraseológicas** y, en general, todo lo que hemos llamado *colocación*.

Ya sabemos que los corpus nos ofrecen la posibilidad de obtener colocaciones para una palabra a partir de la búsqueda de sus colocados a la distancia que queramos (*Corpus del español*) o coapariciones si buscamos en el CORPES. Pero el *Corpus del*

español, además, permite un modo de obtención de colocaciones que no hemos visto hasta ahora. Es a partir de su sección WORD AND PHRASE (<https://www.wordandphrase.info/span/>) que es mucho más intuitiva para que puedan usar nuestros alumnos. En esta sección podemos, también, introducir nuestros propios textos (o los de nuestros alumnos) para analizarlos.

Para hacer una búsqueda hay que clicar en el símbolo de la lista de viñetas. A partir de ahí, la interfaz es parecida a la que estamos acostumbrados, pero la salida es mucho más rica (figura 6). Nos ofrece la frecuencia de la palabra que hemos buscado, en este caso, *aceite*, sus traducciones al inglés, sinónimos, colocados más habituales, palabras/temas relacionados y, abajo, las líneas de concordancias para que nuestros estudiantes puedan ver cómo funciona la palabra en contexto. Además, nos muestra en qué tipo de corpus aparece más, en este caso, SPOK (*spoken*, 'hablado'), FIC (*fiction*, 'ficción'), NEWS ('noticias') y ACAD (textos académicos).

The screenshot shows the 'WordAndPhrase - Spanish' web interface. The top navigation bar includes 'BROWSE LIST', 'LIST OF WORDS', 'SEE ENTRY', and 'ANALYZE TEXT'. The main content area is titled 'ACEITE n (RANK 1750, FREQ 94959)'. It features a progress bar, a 'DEFINITIONS (click for more)' section with a small image of an oil bottle and text 'oil, olive, oil, vegetable oil (bribe)', and a 'MORE: Ling, WR, Rev, Coll, SpD, Oxf' link. Below this is a 'Corrections / problems' section with a question mark icon. The 'COLLOCATES' section shows a bar chart for 'SPOK', 'FIC', 'NEWS', and 'ACAD' corpora, with 'SPOK' being the most frequent. It lists collocates: (ADJ) vegetal, esencial, virgen, extra, caliente, mineral, usado, graso, hidrogenado, seco; (NOUN) oliva, aceite, coco, cucharada, grasa, sal, palma, pescado, almendra, motor; (MISC) freír, mezclar, calentar, ungir, cocinar, extraer, untar, hervir, verter, derramar. The 'SYNONYMS (click to see word)' section lists: grasa, grease, manteca, unto, sebo, aceite, (pringue), lubricante, aceite, graso, lubricante, óleo, (olio) unto, graso, (oleína), lípido, sebo, aceite. The 'TOPICS (click to see)' section lists: grasa, oliva, sal, ingrediente, receta, graso, piel, cocinar, alimento, huevo, azúcar, harina, dieta, cebolla, vitamina, ácido, cucharada, crema, ajo, queso, salsa, sabor, plato, tomate, leche, carne, mezclar, fruta, pescado, verdura.

Figura 6. Ejemplo de búsqueda en la sección WORD AND PHRASE del Corpus del Español

En las líneas de concordancias, como es habitual, también aparece la página web. En este caso en concreto, al hacer la búsqueda se puede comprobar que muchas son páginas de recetas de cocina. Entre las actividades posibles podemos extraer el significado de una lista de vocabulario (o una palabra) por las líneas de

concordancia, como veíamos en el epígrafe anterior, pero también podemos hacer que los alumnos saquen provecho de las colocaciones.

Para ello, una actividad posible sería que, una vez hecha una redacción (por ejemplo, la redacción clásica de preparar una receta de cocina), en la corrección, el profesor marque todas aquellas palabras susceptibles de formar colocaciones, por ejemplo, los ingredientes usados o algunos verbos. Y a partir de ahí el alumno puede buscar las colocaciones de esas palabras para enriquecer su redacción. Por ejemplo, si el alumno ha dicho que usará la palabra *aceite*, puede buscar las colocaciones y elegir si su aceite será *de oliva*, *de oliva virgen* o *de girasol*. Si usa la palabra *cebollas*, puede decidir si son *dulces*, *rojas*, *moradas*, *caramelizadas*...

Otra actividad posible con los resultados de colocaciones nos permitiría trabajar verbos de significado similar, que sin embargo se combinan con nombres diferentes. Sería el caso de los verbos *saber* y *conocer*, ejemplo propuesto por Virginia González García en una entrevista que aparece como material en el apartado 8.7. y que recomiendo encarecidamente a aquellos que quieran tener más ideas para llevar al aula. Dado que es un *podcast*, se puede descargar y aprovechar para escucharlo en el transporte público o en el coche. En las actividades en las que se trabajan verbos de significado similar, se podría trabajar en grupo la búsqueda de colocaciones y después poner en común los resultados para hacer una lista común de toda la clase.

Pasemos a otro tema típico del enfoque léxico, la derivación. En este caso, para buscar en corpus tendremos que usar el asterisco para que nos devuelva palabras que empiecen por el prefijo deseado, pongamos *anti-*, como en *antisocial*, *antivirus* o *anticapitalista*. Para hacerlo, en el corpus del español se puede buscar “*anti**”. El diccionario de la Academia (<https://dle.rae.es/>) también permite realizar búsquedas seleccionando EMPIEZA POR. Hay que tener en cuenta que este tipo de búsquedas no parten de una anotación morfológica, es decir, no podemos decir al corpus: “devuélveme las palabras que contengan el prefijo *anti-*”, sino que nos devolverán todas las palabras que empiecen por *anti-* y en esta categoría entrarían palabras como *antiguo*, donde *anti-* no es un prefijo que significa *en contra*, sino que forma parte de la raíz de la palabra³.

3 El DLE, en la actualidad, solo devuelve los 40 primeros lemas que empiezan por la cadena seleccionada.

En los dos casos, puede ser interesante usar dinámicas de búsqueda en corpus en parejas para facilitar su uso y, después, ponerlo en común en grupos o en clase abierta.

8.5. ADAPTACIÓN DE ACTIVIDADES A CLASE INVERTIDA

La **clase invertida** o *flipped classroom* consiste en que los alumnos realicen un aprendizaje antes de la clase y que la clase sirva para realizar actividades que sirvan para afianzar ese aprendizaje (Milman, 2012). Este sistema es especialmente útil en clases de lengua extranjera cortas, ya que si usamos el tiempo de clase para dar explicaciones gramaticales teóricas y realizar actividades muy formales como vacíos de información, no queda tiempo para practicar con actividades más abiertas y destrezas como la producción oral, que solo se pueden practicar en el aula.

La adaptación del uso de corpus a la clase invertida parte de la base de que los alumnos ya saben buscar en corpus de manera autónoma. Por ello, es recomendable haber realizado alguna de las actividades expuestas en el capítulo anterior dentro del aula. Una vez que los alumnos tengan práctica suficiente con el uso de corpus, se pueden adaptar las actividades para que la parte de búsqueda y los ejercicios de abstracción de la regla gramatical subyacente a los datos se hagan en casa y, a partir de ahí, en la clase, aclarar dudas sobre la regla gramatical, especificar excepciones (si es necesario) y trabajar esos contenidos de manera más abierta.

Hay que tener en cuenta que la búsqueda con corpus desde casa puede no ser apta para todos los públicos. Lo primero que hay que considerar es que se está dando por hecho que los estudiantes tienen un ordenador con conexión a internet disponible. Aunque esto es cierto para la mayoría de ellos, puede no ser tan habitual, por ejemplo, cuando se trabaja en países hispanohablantes con migrantes recién llegados. En segundo lugar, no todo el mundo tiene la misma facilidad con la informática: aunque hayamos trabajado con corpus en clase, para algunos alumnos puede suponer una curva de aprendizaje muy larga que no tiene que ver con la lengua, sino con las herramientas informáticas. Estos alumnos serán muchas veces de mayor edad, cuya motivación para estudiar español es que pasan los periodos vacacionales en países hispanohablantes y no estudiar para un examen. Para ellos el tiempo estará mejor invertido aprendiendo lengua de una manera más comunicativa al uso (por ejemplo, con juegos de rol) que invirtiéndolo en aprender a manejarse realizando búsquedas detalladas en internet.

8.6. LA PRONUNCIACIÓN CON CORPUS

El ámbito de la pronunciación es uno de los que más se puede beneficiar del uso de corpus. Los corpus orales dan acceso a grabaciones que pueden servir para ofrecer al estudiante un modelo diferente al del profesor. Además, estos audios pueden servir para diseñar actividades de entrenamiento de la pronunciación. En internet podemos encontrar varios tipos de recursos que pueden servir para trabajar la pronunciación.

El primer tipo de herramienta disponible, aunque no son exactamente un corpus, son los diccionarios de pronunciación. WordReference ofrece a los usuarios que buscan una palabra, tanto en el diccionario español monolingüe (definición), como en los bilingües la pronunciación de la palabra buscada en tres variedades: España, México y Argentina.

Otro servicio que ofrece características parecidas es la Wikipedia, aunque no para todas sus entradas. En español, casi todas las entradas que incluyen el lema en audio son extranjerismos o nombres propios extranjeros, como Barack Obama. En esos casos, aparece un símbolo de audio en el inicio de la definición. También en Wikipedia existe un proyecto para locutar sus entradas: el índice de entradas locutadas se puede ver aquí (https://es.wikipedia.org/wiki/Wikiproyecto:Wikipedia_grabada). En el caso de las entradas locutadas, el símbolo de audio está en la parte derecha de la línea del lema. Dado que, en Wikipedia, el sonido tiene una licencia Creative Commons, se puede usar gratis con fines educativos. Para bajar el archivo solo hay hacer clic con el botón secundario del ratón sobre el símbolo del sonido o del reproductor.

Estos grandes bancos de datos pueden ser usados por el profesor para construir los clásicos ejercicios de repetición con una voz diferente de la suya sin necesidad de grabarse. Además, puede ser un recurso muy útil para que los alumnos consulten en casa cómo se pronuncia una palabra determinada.

Otro recurso útil para el aula de ELE son los atlas de entonación (Martínez Celdrán & Fernández Planas, 2003; Prieto & Roseano, 2009). La entonación es una de las grandes olvidadas en el aula de ELE. Si quieres saber qué enseñar de ella puedes consultar Elvira-García (2016), pero es, además, uno de los rasgos que está más sujeto a variación dialectal. Por eso puede ser una buena idea enseñar a los alumnos cómo suena una pregunta en Madrid (entonación ascendente) y cómo suena en Cuba (entonación descendente, llamada en la tradición hispánica, circunfleja). Los atlas en línea contienen mapas donde se puede ver, por ejemplo, la entonación de las interrogativas o de los vocativos en diferentes puntos.

En cuanto a los corpus orales propiamente dichos, sus grabaciones pueden servir para ofrecer a los alumnos conversaciones reales (solo en los niveles más altos) o montar juegos de imitación y rol con pasajes escogidos. Una actividad interesante, por el hecho de forzar aquello que llamamos “foco en la forma”, que es esencial en los ejercicios de pronunciación, consiste en usar metraje de películas con subtítulos (V.O. española) para realizar en clase un doblaje de alguna escena de la película. De esta manera, se fuerza a los alumnos a realizar una lectura expresiva del texto, que hará que se pierda, en parte, la monotonía de la entonación de interlengua.

8.7. PARA MÁS INFORMACIÓN

Entrevista a Virginia González García (Universitat de València) en el blog ELE de Lengua “Corpus en la clase de ELE”: <http://eledelengua.com/ldelengua-129-hablando-sobre-corpus-en-el-aula-de-ele/>.

8.8. ACTIVIDADES

8.8.1. Piensa un conjunto de búsquedas que podrían dar juego en un tema dedicado a las profesiones para mostrar la caída y auge de algunas de ellas. Por ejemplo, si buscáramos *pregonero*, veríamos que su uso ha caído. También lo puedes enfocar al género, a partir de búsquedas como *secretario* y *secretaria* (si buscas esta en concreto, te llevarás una sorpresa y verás que no tiene sentido llevarla a clase).

8.8.2. En español existe una construcción de negación que permite combinar el adverbio con la presencia de otro adverbio o indefinido de sentido negativo, palabras como *nunca, jamás, tampoco, nadie, nada o ninguno*. Diseña una actividad de corpus para que los alumnos puedan deducir cuándo una de estas palabras negativas aparece acompañada del adverbio *no*. Es probable que la búsqueda en corpus que les propongas contenga, además de la forma de la palabra de interés, el adverbio *no* como colocado.

8.8.3. Piensa y anota la búsqueda que tendrían que llevar a cabo tus alumnos en el *Corpus del español* para obtener las colocaciones más habituales de *volverse* + adjetivo.

8.8.4. Propón una actividad en la que los estudiantes tienen que deducir el significado de una palabra a partir de las líneas de concordancias.

8.8.5. Entra en el *Atlas interactivo de la entonación del español* y busca la entonación de las preguntas absolutas > de tipo neutro > preguntas informativas de Madrid (España) y Santiago de los Caballeros (República Dominicana). El tono está marcado por la curva negra horizontal ¿cuál te parece la diferencia más sobresaliente entre las dos? Madrid está en el mapa de España (http://prosodia.upf.edu/atlasentonacion/mapa_e.html) y Santiago de los Caballeros en el mapa de América (http://prosodia.upf.edu/atlasentonacion/mapa_l.html).

9

LOS CORPUS PARA LA CORRECCIÓN

La corrección de errores consiste en la rectificación de la desviación de la norma de las producciones de nuestros estudiantes; por lo tanto, siempre que corregimos hay muestras de lengua involucradas (Ferrán, 1990). Es decir, la corrección usa datos. Y, siempre que hay datos, los corpus nos pueden ayudar a ordenarlos de manera más eficaz. Estaremos hablando, por lo tanto, de casos de “haz tu propio corpus” a partir de los datos de tu clase. En realidad, en sentido estricto, esto no serían corpus (a no ser que lleves su creación hasta las últimas consecuencias), pero vamos a aprovechar unas cuantas estrategias de corpus para corregir más sistemáticamente. Además, los alumnos pueden usar los corpus ellos mismos para corregir sus producciones; de ello también hablamos en este capítulo.

9.1. EL USO DE CORPUS PARA LA CORRECCIÓN

En un proceso de corrección normal, los profesores corregimos a partir de nuestro conocimiento de la lengua e intuiciones. Eso hace que muchas veces corrijamos formas lingüísticas porque “esto en español no existe” o “no se dice así”.

Pedro Gras (2018) nos alerta de que esto tiene un peligro: muchas veces podemos acabar diciendo que algo no existe cuando en realidad lo único que pasa es que no está en nuestro **idiolecto**, nosotros no lo decimos. Un ejemplo de esto serían los casos en los que nuestros estudiantes usan formas propias de otras variedades (dialectos), que para nosotros no son gramaticales, pero también dentro de nuestra misma variedad puede haber formas que no se usen en nuestro entorno y, sin embargo, sean normales.

Ante la duda o, incluso, aunque no dudemos, antes de corregir podemos buscar en corpus para asegurarnos de que eso no se dice. Así, antes de corregir por errónea una forma como *¿Me regala la cuenta?*, podemos comprobar en un corpus que esta es la forma habitual de pedir la cuenta en Colombia.

Como vemos esto puede ser útil para corregir tareas entregables, pero también para encontrar un argumento de autoridad en discusiones que se puedan dar en el aula. Hay determinados alumnos que pueden argumentar en clase ante una corrección “pues mi antiguo profesor lo decía así” o “yo lo he oído así”. Tener a mano un corpus con ejemplos reales proporciona pruebas mediante casos de uso.

9.2. CORPUS HISTÓRICO DE PRODUCCIONES DE ALUMNOS

Una práctica habitual en ELE es, después de haber, por ejemplo, corregido redacciones o haber estado oyendo una actividad oral por las mesas, recoger los errores que se han cometido más veces y explicarlos de forma anónima, e, incluso, cambiándolos un poco en clase abierta en la pizarra. Al hacer esto, se está creando un corpus *ad hoc* de la actividad.

La estrategia para crearlo y presentarlo consistiría en recoger todos los errores de la actividad, clasificarlos por el tipo de error, cambiar lo posible para evitar que los alumnos se vean demasiado reflejados o que sus compañeros puedan reconocer su trabajo y dejar que los alumnos los corrijan. En los casos en los que sea necesario, puede corregirlo el profesor o dar una segunda explicación.

Crear tu propio corpus de errores con las clases que trabajes hará que para el año siguiente puedas prever los errores que cometerán. Todos los profesores hacemos esto de manera inconsciente. Si llevas años trabajando en población francófona o italianos, te acostumbrarás a darle importancia a las frases de *cuando* + subjuntivo. Si trabajas con lusófonos, aprenderás con los años a pasar ese tema mucho más deprisa.

Este ejemplo parece obvio y para centrar tu enseñanza en una lengua quizás no necesites un corpus, pero apuntar cada año los errores de tus alumnos y ordenarlos puede centrar tu enseñanza mucho más no solo hacia sus errores, sino también hacia sus intereses. Crear un corpus de tu grupo hará que tengas los errores de tu población (entendida como término estadístico) habitual, es decir, tendrás una muestra con la franja de edad que sueles tener, nivel de instrucción y socioeconómico, etc.

Pero, además, si guardas un registro de las redacciones, tendrás una muestra de los intereses de tus alumnos. Por ejemplo, si en la redacción “biografía de un personaje famoso” todos los años aparece Frida Kahlo, puede ser interesante usar ese personaje para montar un ejercicio de otro tipo, ya que gran parte de los alumnos la conocen y se interesan.

Además, guardar un registro puede aportar un valor adicional. Un vistazo a los temas sobre los que escriben los alumnos te puede indicar cuando los materiales que usas se han quedado desfasados.

Si el manual habla de Michael Jordan (o Fernando Martín), pero vemos que las redacciones de tus estudiantes han pasado por Kobe Bryant, Pau Gasol y han llegado a Ricky Rubio y Alba Torrens, quizás sea el momento de plantearse cambiar el manual por uno actualizado. No necesitamos saber de baloncesto, solo prestar atención a los jugadores preferidos de nuestros alumnos.

Estos corpus creados por nosotros con las producciones de nuestros alumnos tienen una particularidad de la que ya hemos hablado. Estrictamente, no serían corpus, al menos no corpus completos, puesto que lo más probable es que para nuestros propósitos no sea necesario hacer una anotación morfológica completa o anotar y clasificar todos los errores, y eso hace que el corpus no sea reutilizable, una característica que Torruella & Llisterri (1999) consideran esencial en la creación de corpus. No obstante, estaremos usando las técnicas de la lingüística de corpus para cumplir nuestros objetivos y eso es lo mejor que podemos hacer por una disciplina, sacarla de lo teórico y hacerla útil.

9.3. USO DE CORPUS PARA LA AUTOCORRECCIÓN

En el mundo del ELE es habitual el uso de la doble corrección con propósitos didácticos. En ella, el profesor realiza una primera revisión de las producciones escritas de los alumnos en las que marca los errores que el alumno debería conocer, por ejemplo, con un subrayado, y permite que sea el alumno quien busque la forma correcta. Una vez que el alumno ha corregido sus errores (o lo ha intentado), el profesor vuelve a corregir la redacción esta vez con el método clásico, proporcionando las formas correctas.

En este tipo de actividades, los alumnos que previamente hayan descubierto los corpus con el profesor pueden usarlos para verificar sus respuestas o incluso para buscar la respuesta correcta. Pongamos que uno de los errores de la redacción marcado por el profesor es *la gente son*, el alumno puede buscar en el CdE (BYU) la palabra *gente* seguida de verbo “gente [v*]” y verá que los dos primeros resultados son *gente es* y *gente está*, seguidos de otros también muy habituales como *gente tiene*, *quiere*, *piensa* o *puede*, pero todos ellos en singular. A partir de ahí, puede darse cuenta de su error y, si tiene dudas de por qué *gente* se combina con singular, acudir a sus propios apuntes, manuales o al profesor.

Además, el uso de corpus para la corrección, en niveles avanzados, también posibilita que los estudiantes puedan corregir aquellas cosas que no se han visto de manera literal en clase. Así, si un alumno escribe en una redacción *lo miró por la esquina del ojo* y lo marcamos como incorrecto, él mismo puede buscar el nombre que va normalmente con *del ojo*, mediante la siguiente expresión “[N*] del ojo”. En este caso, habría que buscar, por ejemplo, en la sección GENRE/HISTORICAL del CdE, ya que esta es una expresión que se usa sobre todo en narrativa y en otras subsecciones del corpus, como el NOW de noticias, no se encuentra. En los corpus que incluyen narrativa, sin embargo, la colocación *raballo del ojo* suele aparecer la primera seguida de otras más técnicas como *retina del ojo* o *fondo del ojo*. Además, el estudiante podrá ver que *raño del ojo* también es posible, aunque mucho menos habitual.

La dificultad de las búsquedas que deberá realizar el alumno dependerá mucho del error. En algunos casos, con buscar la palabra directamente se encontrará con la solución, pero los errores más relacionados con cuestiones contextuales requieren búsquedas más complejas. Por eso, en algunos casos puede ser necesario que el profesor ofrezca al estudiante la expresión que debe buscar en el corpus. Siguiendo con el ejemplo de arriba, en el caso de *la gente son* podríamos subrayar la palabra *son* e indicar en la corrección “gente [v*]” para que el alumno lo pueda buscar. Si todos los errores son del mismo tipo, con ofrecer un ejemplo será suficiente. Por supuesto, esto dependerá de cuánto se hayan trabajado los corpus en clase. Si los alumnos ya son expertos en el uso del CdE, no será necesario y los estudiantes pueden hacer las búsquedas probando a ensayo y error de manera autónoma.

Este tipo de actividades incluidas en tareas de corrección fomentan el aprendizaje por descubrimiento y ayudan a los alumnos a prestar atención a algo que en algunos métodos pasa desapercibido y es que la lengua no es absolutamente modular (no es siempre la suma de sus partes). Aunque podamos crear significados nuevos mediante la modularidad (puedo añadir *nuevo* o *viejo* a la palabra *coche* para modificarlo), la lengua que hablamos tiene un alto porcentaje de composicionalidad (es más habitual decir *coche nuevo* que *coche flamante*) y, por lo tanto, las palabras suelen aparecer en los textos y en el habla en compañía. Estudiarlas aisladas y después escribirlas así, crea un tipo de lengua no nativa, mientras que usar bien las colocaciones hace que las producciones se parezcan más a las de los nativos (García Salido & García, 2018).

9.4. PARA MÁS INFORMACIÓN

Ferrán, J. M. (1990). La corrección del error: fundamentos, criterios, técnicas. En: Bello, P. et ál. (eds.). *Didáctica de las segundas lenguas. Estrategias y recursos básicos*. Madrid: Santillana.

9.5. ACTIVIDADES

9.5.1. Imagina que un alumno ha cometido los siguientes errores. Indícale las búsquedas que debería realizar y en qué corpus (CdE o usar la web como corpus mediante Google) para encontrar las formas correctas. Por ejemplo, *esquina del ojo > CdE, busca “[N*] del ojo”. La referencia al final de los ejemplos indica el corpus de dónde se han extraído.

- *La noche* doy una vuelta por la ciudad. (Boudali, 2006, 1.º nivel, R1).
- He *asistido a* muchos espectáculos del festival del Jazz (Boudali, 2006: 1.º nivel, R9).
- Yo creía que ella era una persona muy importante *para* mi vida (Gutiérrez, 2005: informante 39, curso 4.º, texto 1).
- A mediodía regresamos a casa para almorzar y *hacer* la siesta (Boudali, 2006: 3.º nivel, R3).

10

CORPUS PARA EL APRENDIZAJE AUTÓNOMO DEL ALUMNO

Viendo el epígrafe de este apartado se puede pensar: ¿Qué hace esto aquí? Si es aprendizaje autónomo, el profesor no tiene que hacer nada, ¿no? Y nada más lejos de la realidad. Incluso los alumnos que prefieren el aprendizaje autónomo agradecerán algún tipo de incentivo para trabajar. Así que, aunque seguramente muchos alumnos ya están haciendo uso de corpus en su casa sin que el profesor intervenga, este capítulo es para a) aquellos alumnos menos acostumbrados al uso de internet que no hayan descubierto estas herramientas y b) obtener un mayor rendimiento en esos alumnos que ya usan corpus en su rutina de escritura.

El uso autónomo de corpus más habitual que realizan los estudiantes es durante la creación de redacciones. La entrada a los corpus al escribir puede ser por diferentes vías. La más directa sucede cuando buscan en Google para comprobar si una palabra existe (por ejemplo, una forma verbal flexionada) introduciéndola tal y como ellos la habían escrito y el buscador les devuelve, o bien ejemplos de corpus, o bien directamente la interfaz de un corpus en línea.

Otra búsqueda habitual es la que hacen para ver cómo se traduce una palabra de su L1 a la lengua meta. Una búsqueda como *How do you say accurate in Spanish* remite en primer lugar a corpus paralelos como Linguee y, en segundo, a herramientas con grandes listas de ejemplos de uso en contexto como la web Bab.la. Otros alumnos más acostumbrados a hacer búsquedas refinadas puede que busquen en Google como expresión literal (usando entrecomillado) un verbo con las preposiciones entre las que dudan, como “consta de” y “consta en”. Y, de esta manera, Google les devolverá el número de casos obtenidos con cada una de las preposiciones, en un caso unos 56 millones y en el otro unos 8 millones; es fácil inferir cuál es la colocación. Por supuesto, “constar en”, también existe como colocación en expresiones como *constar en acta*, pero es mucho menos habitual. Cuando los alumnos buscan en Google, sea con una búsqueda simple o avanzada, están usando la web como corpus.

Pero en todos los casos, los alumnos acuden a estos corpus en momentos concretos y normalmente a partir de una búsqueda sencilla en Google, sin saber lo que es un corpus ni las posibilidades que ofrece. Y acaban consultando, por ejemplo, un corpus paralelo de los textos de EuroParl (que contiene los discursos del parlamento europeo traducidos a todas las lenguas oficiales de la Unión). Esto significa que, aunque están haciendo uso de las herramientas de corpus, probablemente no le están sacando todo el partido que podrían.

Buyse & Verlinde (2013) realizaron un estudio sobre el uso que hacían sus estudiantes de estas herramientas. Para ello, seleccionaron dos grupos de estudiantes. De ellos, un grupo realizaría las redacciones usando estas tecnologías y el otro grupo no. El resultado fue que el grupo que usó ayuda de corpus en línea realizó redacciones con más riqueza léxica y un mayor nivel de corrección, sin embargo, invirtió más tiempo en escribirlas. Este mayor tiempo se puede deber a dos motivos: el primero, al tener la posibilidad de comprobar si lo que se ha escrito es correcto, se puede producir una sobreconsulta y buscar cosas de las que en realidad ya estaban seguros y, el segundo, el uso de corpus a veces no es todo lo intuitivo que nos gustaría y acostumbrarse lleva tiempo. En cualquier caso, el aumento de tiempo no es algo exclusivo de los corpus.

En los tiempos en los que no existía el corrector automático y consultábamos las palabras en un diccionario, realizar una redacción en la que se habían consultado las palabras también era más lento que entregarla sin hacerlo y, por lo tanto, con faltas de ortografía. La cuestión aquí es cómo se aprende más. Y los datos apuntan a que los alumnos retienen más las grafías y las colocaciones que han tenido que buscar en un diccionario o en un corpus a aquellas que simplemente han visto corregidas por el profesor en un momento en el que están más preocupados por el resultado de la corrección, ya sea una nota o sea el “lo he hecho bien” (cuánto rojo hay en el papel) que en centrarse en los errores.

En cualquier caso, para paliar la inversión de tiempo de nuestros estudiantes en el uso de estas herramientas, Buyse & Verlinde (2013) proponen que haya un papel activo del profesor para ayudar a los estudiantes a que sean más eficientes usando las herramientas. Cómo conseguir acostumbrar a los alumnos y a los profesores a la consulta regular de corpus es una cuestión que todavía queda por resolver. Por eso, vamos a dedicar nuestro próximo y último capítulo a las secuencias didácticas.

10.1. PARA MÁS INFORMACIÓN

Buyse, K. y Verlinde, S. (2013). Possible Effects of Free on Line Data Driven Lexicographic Instruments on Foreign Language Learning: The Case of Linguee and the Interactive Language Toolbox. *Procedia - Social and Behavioral Sciences*, 95, 507–512. <https://doi.org/10.1016/j.sbspro.2013.10.675>.

10.2. ACTIVIDADES

10.2.1. Piensa una actividad en la que tus alumnos tengan que usar recursos de corpus para hacer una redacción. Pon especial cuidado en poner límites para que no busquen cada una de las palabras y colocaciones de la redacción.

11

SECUENCIAS DIDÁCTICAS DE CORPUS Y PROGRESIÓN DEL USO DE CORPUS EN UN CURSO

Para acabar este manual, parece adecuado ofrecer una visión global de cómo usar los corpus en clase. El objetivo es que tengamos una visión integradora de los corpus con ejemplos de cómo usar los corpus en cada una de las partes de la clase. Para ello, empezaremos por la contextualización de la clase, resumiremos cómo usar corpus en una secuencia didáctica completa, veremos una lista de cosas que hay que comprobar cuando se van a usar corpus y, por último, veremos un ejemplo de secuencia didáctica.

11.1. CONTEXTUALIZACIÓN

Ya sea para una secuencia didáctica (un conjunto de actividades que se verán en un día) o para el uso de corpus en un curso, las actividades que hemos visto no se pueden presentar al alumno directamente (puede que huyan despavoridos). Los corpus, como toda la tecnología, tienen una curva de aprendizaje y, para que los alumnos estén dispuestos a asumirla, tienen que sentir que merece la pena, que van a obtener un beneficio de aprendizaje o, por ejemplo, una mejora en sus redacciones.

Cuando vamos a usar los corpus en clase hay dos obstáculos que tenemos que superar. El primero es que las interfaces no suelen ser agradables y el segundo es que muchos alumnos tienen reticencias con la tecnología en general. Como hemos dicho, para superarlas tenemos que mostrarles que los corpus valen la pena. Y, para ello, ¿qué mejor que mostrarles su utilidad en clase?

La mejor introducción que podemos hacer a los corpus es que los alumnos nos vean usarlos en clase, por ejemplo, para resolver sus dudas. Podemos empezar aplicando la búsqueda de ejemplos en clase; de esta manera, los alumnos verán en pantalla los resultados de la búsqueda de corpus, que nosotros explicaremos, y la interfaz les empezará a resultar conocida. Al poco tiempo, quizá ellos mismos se empiecen a interesar y a usarlos.

En el caso de la preparación de una secuencia didáctica, el mismo objetivo se puede alcanzar a partir de la contextualización de la tarea. En esa contextualización, podemos mostrar y explicar qué es un corpus y realizar búsquedas que puedan resultar divertidas para nuestros estudiantes. Y probablemente, el más simpático, con una interfaz más bonita y fácil de usar y de entender es Google Ngram Viewer. Pero tenemos otras opciones para que nuestros estudiantes abran boca. Una manera de motivar una búsqueda en corpus podría ser:

1. Buscar *te quiero un* en Google y enseñar lo que pone en el autocompletado;
2. los resultados serán *te quiero un poco, un montón, un mundo, un potosí, un huevo*;
3. preguntar a los estudiantes en clase abierta cuál creen que es más habitual y por qué;
4. y, para descubrir cuál es más frecuente, abrir con ellos un corpus y ordenar las colocaciones por frecuencia.

De manera ideal, en la contextualización mostraremos ejemplos del corpus que van a tener que usar los alumnos. De esta manera, habrán visto cómo se busca en él y nos les resultará tan extraño. Entonces, pueden empezar las actividades.

11.2. CONTENIDO DE LA CLASE

Normalmente, en una secuencia didáctica integrada tenemos una parte de trabajo del léxico y de la gramática, y las actividades van de las más formales a las más abiertas. Por ello, en la primera parte de una buena secuencia didáctica se podría emplear corpus para buscar colocaciones a partir del léxico que aparece en un documento de texto o vídeo (que podría estar también extraído de un corpus). En esta parte, se pueden usar o adaptar algunos de los ejemplos que se han visto en la sección de actividades de enfoque léxico (8.4).

Después, pasaríamos a la gramática donde haríamos a los alumnos reflexionar sobre el uso (primero) y la forma (después) del ítem gramatical que queremos enseñar. En esta sección se podrían adaptar los ejemplos de 8.3 y las actividades que tenían que ver con la gramática. De manera paralela, si en la parte de estudio de la forma lingüística preparamos ejercicios de vacíos de información, los ejemplos también podrían proceder de corpus.

Por último, pasaríamos a actividades más abiertas que podrían empezar con actividades en las que los alumnos tienen que usar las colocaciones o tiempos verbales aprendidos en contexto para crear sus propias frases. Y, por último, realizar una tarea de escritura o una tarea oral abierta, pero de un tema relevante para la secuencia.

Por supuesto, en una clase al uso no hay por qué tocar todos los palos; algunas clases se enfocarán más a la enseñanza del léxico y otras a la gramática. Esta es solo una propuesta de cómo aunar todo en una misma secuencia. Pero como siempre, lo más importante es que la clase sea coherente y que cumpla los objetivos de aprendizaje del tema. Con la misma filosofía, antes de diseñar las actividades hay que pensar en si el uso de corpus va a ser relevante para nuestros propósitos y si las búsquedas que se van a realizar son clarificadoras para los alumnos. Si pensamos que los ejemplos que van a obtener en los corpus van a conseguir confundirlos y no ayudarlos, no tiene sentido que incluyamos corpus en las actividades. Por eso, es interesante esta lista de requisitos que podemos comprobar al plantear actividades con corpus (esquema 7).

¿Qué nos debemos plantear en nuestra secuencia didáctica con corpus?

- Hemos contextualizado y presentado qué es un corpus.
- El uso de corpus tiene sentido, aporta valor a la actividad.
- El corpus que se usa es adecuado al objetivo de la actividad.
- El resultado de la búsqueda es una ejemplificación clara del fenómeno que queremos ilustrar.
- Las dinámicas (individual, pareja, grupo, clase abierta...) son adecuadas a la búsqueda.

Esquema 7. Lista de requisitos en el diseño de actividades con corpus

11.3. EJEMPLO DE SECUENCIA DIDÁCTICA

En esta secuencia didáctica, adecuada para alumnos de nivel A2, vamos a acometer la enseñanza de los tiempos de pasado mediante el uso de la biografía o historia de vida. Este planteamiento didáctico es el habitual para los tiempos de pasado en A2 y se incluye en la mayoría de los manuales.

En concreto, nos vamos a centrar en la enseñanza del indefinido (también llamado pretérito perfecto simple, como la forma *canté*) en formas de tercera persona, para después pasar a formas de primera y segunda, siempre en singular. El léxico que se incluye son los verbos relacionados con experiencias vivenciales y sus regímenes de preposición (*vivir en, vivir con...*).

El tiempo previsto para la realización de la totalidad de la secuencia es de dos horas (hay que tener en cuenta que en clases de dos horas suele haber una pausa de 10-15 minutos). Los enunciados se han redactado en su mayoría en presente de la primera persona del plural (*nosotros hacemos*) porque el imperativo se introduce, según el PCIC, al final de un A2 (aunque es cierto que en muchos cursos se introduce antes).

El corpus usado es *Corpus del español* de Davies (CdE). Se ha optado por este corpus porque, pese a tener unas búsquedas complejas, permite obtener las colocaciones ordenadas por frecuencia de manera directa, mientras que, en otros corpus, como en el CORPES, es necesario ordenarlas usando los filtros lo que hace que queden en primera posición elementos no relevantes tales como signos de puntuación.

La secuencia empieza haciendo uso de un recurso de clase invertida: una búsqueda en internet que se realiza en casa y que sirve para introducir el tema vertebrador de la unidad (las historias de vida o biografías) y el contenido gramatical (tiempos de pasado, el indefinido) mediante una cita que seguramente los alumnos conocerán en su lengua y que ofrece tres indefinidos seguidos. Además, la actividad sirve como introducción al uso de la tecnología y las búsquedas de lengua en internet; de una manera laxa, los estudiantes están usando la web como corpus, lo que ayudará a ver el sentido de las siguientes actividades.

La secuencia sigue con una pequeña comprensión lectora que se usará para introducir el contenido léxico (verbos de experiencias en biografías y las preposiciones con las que se combinan) y el contenido gramatical del tema (tiempos de pasado, el indefinido). Las siguientes dos actividades usan de manera implícita la noción de colocación y líneas de concordancia. En ellas se trabajan las preposiciones con las que se combinan los verbos que aparecen en el texto a través de la búsqueda en el propio texto y en corpus. Las dinámicas más adecuadas en este caso son de pareja, dada la dificultad añadida de la búsqueda en corpus. Seguidamente, se trabaja el uso y forma del indefinido también a través de la búsqueda en corpus y en parejas. Y se finaliza la clase con ejercicios más abiertos de creación de frases y diálogo con lo que se da una progresión desde actividades más formales, con vacíos de información, hasta diálogos más abiertos (aunque con un modelo de lengua claro, ya que estamos en A2). En las actividades, se indican entre paréntesis anotaciones para el profesor.

VINO, VIO Y VENCÍO

Vida de los famosos

La biografía y el pasado

 **Buscamos en internet *Vine, vi y vencí*. ¿De quién es la frase? ¿Cómo se traduce en tu lengua?** (*Inicio con clase invertida, en el aula, comentar las traducciones a sus lenguas y la original en latín de Veni, vidi, vici.*)

1. Leemos el texto y contestamos las preguntas.

César fue un militar y político romano. Nació en Roma, pero desde joven vivió en muchos lugares, como en Bitinia. Estudió astronomía y latín en casa y escribió muchos libros. Fue a muchos lugares por trabajo, por ejemplo, viajó a Hispania. También, trabajó en la Galia, que conquistó rápidamente. Lo asesinaron con 56 años. (*El texto está adaptado de Wikipedia.*)

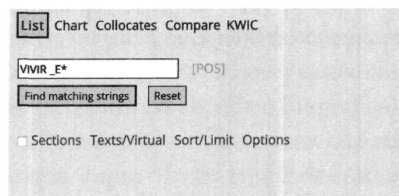
- ¿De qué trabajó César?
Fue militar (general romano) y político.
- ¿Dónde nació?
En Roma.
- ¿Qué estudió César?
Astronomía y latín.
- ¿Cómo conquistó la Galia?
Muy rápido.
- ¿Cómo murió?
Lo asesinaron / mataron.

2. Rellenamos los huecos. ¿Con qué preposiciones se combinan los siguientes verbos en el texto? (*La actividad se puede realizar de manera individual o en parejas.*)

- Nacer *en* un lugar.
- Vivir *en* un lugar.
- Viajar *a* un lugar.
- Trabajar *en* un lugar.
- Ir *a* un lugar.

3. Ahora, ampliamos la lista. En parejas, buscamos otra preposición con la que también se combinan estos verbos y anotamos con qué palabras se usan y un ejemplo. Para hacerlo, vamos a usar una herramienta nueva, los corpus. Seguimos las instrucciones para encontrar las palabras. (El profesor debe enseñar el corpus y realizar la búsqueda de ejemplo en clase abierta para facilitar el trabajo a los alumnos.)

- Vamos a <https://www.corpusdelespanol.org/web-dial/>.
- Escribimos el verbo, por ejemplo, "VIVIR" (en mayúscula para buscar todas las formas conjugadas) y en colocados "_E*" (busca solo preposiciones), como en la imagen, y pulsamos BUSCAR.



- Tenemos la lista de preposiciones con las que se combina. Apuntamos la más frecuente. (No puede ser una de las que salen en el ejercicio anterior.)
- Pulsamos para ver ejemplos de frases, así vemos para qué se usan.

VERBO	PREPOSICIÓN	PALABRA CON QUE SE COMBINA	EJEMPLO
vivir	con	alguien	vivir con la pareja
viajar	en	medio de transporte	viajar en tren
trabajar	con/para	alguien / una empresa	trabajar para Zara
ir	en	medio de transporte	ir en avión

4. En el texto, hay un tiempo verbal nuevo. Buscamos en el corpus "VIVIR_VIS*", para tener una lista de ese tiempo. Pulsamos en el primero (vivió) para ver ejemplos. ¿Para qué se usa *vivió*?

- Para hablar del pasado ☒
- Para hablar del presente ☐
- Para hablar del futuro ☐

5. Rellenamos la siguiente tabla. Para hacerlo, buscamos en el corpus "VIVIR_VIS*". ¿No sabes de qué persona es cada forma? Mira los ejemplos. (La primera persona del plural no aparece en el corpus, por eso la proporcionamos como ejemplo.)

	VIVIR
yo	viví
tú	viviste
él - ella - usted	vivió
nosotros	vivimos
vosotros	vivisteis
ellos - ellas - ustedes	vivieron

6. Creamos frases sobre nuestra vida.

- Nací en ...
-
-

7. Descubrimos la vida de nuestro compañero. Preguntamos dónde nació, *vivió* y *viajó*. Después, se lo explicamos a una tercera persona.

- ¿Dónde naciste?
- Nací en Berlín

11.4. PARA MÁS INFORMACIÓN

Martín Peris, E. (1996). *Las actividades de aprendizaje en los manuales de E/LE*. Tesis doctoral. <https://www.tdx.cat/handle/10803/286731>

11.5. ACTIVIDADES

11.5.1. Plantea una secuencia didáctica para tratar el uso y forma del subjuntivo. No es necesario que la desarrolles en su totalidad, solo apunta los temas que tratarías y qué corpus usarías y por qué.

BIBLIOGRAFÍA

- Abercrombie, D.** (1965). *Studies in Phonetics and Linguistics*. Oxford University Press.
- Ainciburu, M. C.** (2012). Recursos para estudiar el español académico a partir de corpus. *Revista Nebrija de Lingüística Aplicada a La Enseñanza de Lenguas*, 12, 209-215.
- Alameda, J. R., & Cuetos, F.** (1995). *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Servicio de publicaciones de la Universidad de Oviedo.
- Alvar, M.** (1969). *Estructuralismo, geografía lingüística y dialectología actual*. Gredos.
- Andión Herrero, M. A., González Sánchez, M., & San Mateo Valdehita, A.** (2019). *¿Qué debe saber un profesor de ELE L2? Guía básica para actuales y futuros profesores*. UNED.
- Anthony, L.** (2019). Waseda University. *AntConc* (versión 3.5.8). Waseda University. <https://www.laurenceanthony.net/software>
- Billières, M.** (2015). Méthode verbo tonale: origine et fondements. *Au son du FLE*. Toulouse. <http://www.verbotonale-phonetique.com/origines-fondements/>
- Blum-kulka, S.** (1980). Learning to say what you mean in a second language: A study of the speech act performance of learners of hebrew as a second language. *Applied Linguistics*, 3(1), 29-59. <https://doi.org/10.1093/applin/III.1.29>
- Bosque, I. (Ed.).** (2004). *Redes. Diccionario combinatorio del español contemporáneo*. SM.
- Bosque, I. (Ed.).** (2006). *Diccionario combinatorio práctico del español contemporáneo*. SM.
- Briz Gómez, A., València, U. de, & Almería, U. de.** (2002). *Corpus de conversaciones coloquiales*. Arco/Libros.

Brown, J. D., & Rodgers, T. S. (2002). *Doing second language research: An introduction to the theory and practice of second language research for graduate/master's students in TESOL and applied linguistics, and others*. Oxford University Press.

Brumfit, C. (1984). *Communicative Methodology in Language Teaching: the Roles of Fluency and Accuracy*. Cambridge University Press.

Buyse, K., & González Melón, E. (2012). El corpus de aprendientes Aprescrilov y su utilidad para la didáctica de ELE en la Bélgica multilingüe. *Multilingüismo y enseñanza de ELE en contextos multiculturales*, 247-261.

Buyse, K., & Verlinde, S. (2013). Possible Effects of Free on Line Data Driven Lexicographic Instruments on Foreign Language Learning: The Case of Linguee and the Interactive Language Toolbox. *Procedia - Social and Behavioral Sciences*, 95, 507-512. <https://doi.org/10.1016/j.sbspro.2013.10.675>

Caballero, M., Díaz, L., & Taulé, M. (2012). Cómo prever y organizar los contenidos lingüísticos y sociopragmáticos de un escenario comunicativo en clase de E/LE a partir del uso de corpus. *Monografías MarcoELE: Estudios Sobre Aprendizaje y Adquisición de Lenguas y Sus Contextos*, 15, 6-26.

Caballero, M., Díaz, L., & Taulé, M. (2014). *Guía de anotación del corpus FerroviELE. Working paper 2: DIANA-Construcciones*. Universitat de Barcelona.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47. <https://doi.org/10.1093/applin/1.1.1>

Cestero, A. M., Penadés, I., Blanco, A., Camargo, L., & Simón, J. (2002). Corpus para el análisis de errores de aprendientes de E/LE (CORANE). *Tecnologías de La Información y de Las Comunicaciones En La Enseñanza de ELE*. En *Actas Del XII Congreso Internacional de la Asociación Para La Enseñanza Del Español Como Lengua Extranjera, 2000* (pp.527-534). Universitat Politècnica de València.

Chomsky, N. (2004). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics*, 1, 93-111.

Cordella, M. (2002). La interacción médico-paciente en escrutinio: un estudio de sociolingüística interaccional. *Onomázein*, 7, 117-144.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Peter Lang.

Coulthard, M. (2013). On the use of corpora in the analysis of forensic texts. *International Journal of Speech Language and the Law*, 1(1), 27-43.

Cruz Piñol, M. (2012). *Lingüística de corpus y enseñanza del español como 2/L*. Arco/Libros.

Crystal, D. (2013). *Clinical linguistics*. Springer.

Cuenca, M. J., & Hilferty, J. (1999). *Introducción a la lingüística cognitiva*. Ariel.

Davies, M. (2002). *Corpus del Español: 100 million words, 1200s-1900s*. <http://www.corpusdelespanol.org>

Elvira-García, W. (2016). La entonación en el aula de ELE. In N. Baulenas & B. Montmany (Eds.), *Dossier del XXV encuentro práctico de profesores de ELE* (pp. 41-47). Difusión.

Ferrán, J. M. (1990). La corrección del error: fundamentos, criterios, técnicas. En P. Bello et ál (Eds.), *Didáctica de las segundas lenguas. Estrategias y recursos básicos*. Santillana.

Ferreira, A., Elejalde, J., & Vine, A. (2014). Análisis de errores asistido por computador basado en un corpus de aprendientes de español como lengua extranjera. *Signos*, 47(86), 385-411.

Fillmore, C. (1991). "Corpus linguistics" or "Computer-aided armchair linguistics." En *Directions in Corpus Linguistics* (pp. 35-66). Mouton de Gruyter.

Fornaciari, T., & Poesio, M. (2012). DeCour: a corpus of Deceptive statements in Italian COURts. *LREC* (pp. 585-1590).

Fornaciari, T., & Poesio, M. (2014). Identifying fake Amazon reviews as learning from crowds. En *14th Conference of the European Chapter of the Association for Computational Linguistics* (279-287).

García Gutiérrez, M. (2005). Corpus de conversaciones en español como lengua extranjera. *LinRed: Lingüística En La Red*, 3.

García Salido, M., & Garcia, M. (2018). Comparing learners' and native speakers' use of collocations in written Spanish. *International Review of Applied Linguistics in Language Teaching*, 56(4), 401-426. <https://doi.org/10.1515/iral-2016-0103>

González Calvo, J. M. (2000). Sobre la palabra y las clases de palabras. *Revista Española de Lingüística*, 30(2), 309-330.

Granger, S. (2003). The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), 538-546.

Gras, P. (2018). Más allá de las gramáticas y los diccionarios, los corpus como herramientas en la enseñanza / aprendizaje de lenguas. *MarcoELE. Revista de Didáctica Español Lengua Extranjera*, 26.

- Gutiérrez Quintana, E. (2005). Corpus de textos escritos por universitarios italianos estudiantes de ELE. *Linred: Lingüística En La Red*, 3. <http://hdl.handle.net/10017/24626>
- Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. Edward Arnold.
- Hardie, A. (2014). Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal*, 38(1), 73-103. <https://doi.org/10.2478/icame-2014-0004>
- Hincapié, D. (2018). Corpus de aprendientes de español como lengua extranjera y segunda lengua (CÁELE/2): el componente escrito. *Forma y Función*, 31(2), 29-143.
- Hong, L., Convertino, G., & Chi, E. H. (2011). Language matters in Twitter: A large scale study. *Fifth International AAAI Conference on Weblogs and Social Media*, 518-521. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPaper/2856>
- Hornby, A. S., Gatenby, E. V., & Wakefield, H. (1948). *The Advanced Learner's Dictionary of Current English* (1956th ed.). Oxford University Press.
- Ide, N., & Macleod, C. (2001). The American National Corpus: A standardized resource of American English. *Proceedings of Corpus Linguistics* (Vol. 3, pp. 1-7). Computer Science Department, New York University
- Ide, N., & Véronis, J. (Eds.). (1995). *Text encoding initiative: Background and contexts*. Springer Science & Business Media.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1).
- Kita, K., Kato, Y., Omoto, T., & Yano, Y. (1994). A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria. *Journal of Natural Language Processing*, 1(1), 21-33. <https://doi.org/10.5715/jnlp.1.21>
- Koerner, E. K. (2003). *Toward a history of American Linguistics*. Routledge.
- Lewis, M. (1993). *The Lexical Approach: The state of ELT and a way forward*. Language Teaching Publications.
- Liberman, M. (2006). *Sex-Linked Lexical Budgets*. Language Log. *Language Log*. University of Pennsylvania.
- Llanos, L. C., Gómez, P. G., & Sandoval, A. M. (2007). El corpus C-ORAL-ROM en la enseñanza de ELE. En Las destrezas orales en la enseñanza del español L2-LE. En *XVII Congreso Internacional de la Asociación del Español como lengua extranjera (ASELE)* (pp. 1115-1128). Universidad de la Rioja.
- López Santiago, M., & Giménez Folqués, D. (2017). *El léxico del discurso turístico 2.0*. Servicio de Publicaciones, Universidad de Murcia.
- Lozano, C. (2009). CEDEL2: Corpus Escrito del Español L2. En C. M. Bretones Callejas (Ed.), *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente* (pp. 197-212). Universidad de Almería. <http://hdl.handle.net/10481/22177>
- Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and SLA: the design and collection of CEDEL2. En N. B. & P.T.A. Díaz-Negrillo (Ed.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 65-100). John Benjamins.
- Martínez Celdrán, E., & Fernández Planas, A. M. (2003). *Atlas Multimèdia de la Prosòdia de l'Espani Romànic*. Universitat de Barcelona. http://stel.ub.edu/labfon/amper/cast/index_ampercat.html
- Matte Bon, F. (1995). *Gramática comunicativa del español II. De la idea a la lengua*. Edelsa. http://catalog.ub.edu/record=b1503132~S1*spi
- McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., & McCullough, D. (2012). On building a reusable Twitter corpus. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1113-1114).
- McEnery, T., & Wilson, A. (1996). En *Corpus Linguistics*. Edinburgh University Press.
- Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are Women Really More Talkative Than Men? *Science*, 317(5834), 82-82. <https://doi.org/10.1126/science.1139940>
- Melero, P. (2000). *Métodos y enfoques en la enseñanza/aprendizaje del español como lengua extranjera*. Edelsa.
- Mighetto, D. (2003). Una aproximación cuantitativa al léxico del español escrito en SOL-Spanish Online. En F. Sánchez Miret (Ed.), *Actas del XXIII Congreso Internacional de Lingüística y Filología Románica: Volume III : Semántica léxica, lexicología y onomástica* (pp. 319-334). Max Niemeyer Verlag.
- Milman, N. B. (2012). The flipped classroom strategy: What is it and how can it best be used? *Distance Learning*, 9(3), 85.
- Moliner, M. (1966). *Diccionario de uso del español* (Vol. 2). Gredos.
- Nicolás Martínez, C. (2012). *C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente)*. Liceus.
- Nicolás Martínez, C. (2015). Las funciones comunicativas en la lengua oral: validez y adecuación al alumnado. En *La enseñanza de ELE centrada en el alumno* (pp. 1197-1208). Asociación para la Enseñanza del Español como Lengua Extranjera.

Nitti, J. J. (1979). Computers and the old spanish dictionary. En *Medieval Studies and the Computer* (pp. 43-52). Pergamon.

Pérez Serrano, M. (2014). Análisis de errores colocacionales en un corpus de aprendientes de ELE. *MarcoELE: Revista de Didáctica Español Lengua Extranjera*, 19, 11.

Pitkowski, E. F., & Gamarra, J. V. (2009). El uso de los corpus lingüísticos como herramienta pedagógica para la enseñanza y aprendizaje de ELE. *Tinkuy: Boletín de Investigación y Debate*, 11, 31-51.

Prieto, P., & Roseano, P. (Eds.). (2009). *Atlas interactivo de la entonación del español*. Universitat Pompeu Fabra.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Real Academia Española. (2014). *Diccionario de la lengua española* (23.ª ed.). Real Academia Española. <http://dle.rae.es/>

Reese, S., Boleda, G., Cuadros, M., & Rigau, G. (2010). Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* 1418-1421. European Language Resources Association (ELRA).

Rojo, G. (2009). Sobre la construcción de diccionarios basados en corpus. *Tradumàtica: Traducció i Tecnologies de La Informació i La Comunicació*, 7, 1-7.

Rojo, G. (2016). Los corpus textuales del español. En J. Gutiérrez-Rexach (Ed.), *Enciclopedia lingüística hispánica* (pp. 285-296). Routledge.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching*, 10(1-4), 209-232.

Sinclair, J. (1987). *Collins COBUILD English language dictionary*. Harper Collins Publishers.

Taylor, C. (2008). What is corpus linguistics? What the data says. *ICAME Journal*, 32, 179-200.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins.

Torruella, J., & Llisterri, J. (1999). Diseño de corpus textuales y orales. *Filología e Informàtica. Nuevas Tecnologías En Los Estudios Filológicos*, 45-77.

Valverde, P. (2020). Diseño y creación de un corpus de aprendientes de español en japon (CELEN). *E-AESLA*, 6, 223-240.

Vangehuchten, L. (2004). El uso de la estadística en la didáctica de las lenguas extranjeras con fines específicos: descripción del proceso de selección del léxico típico del discurso económico empresarial en español. En *7 Es Journées Internationales d'Analyse Statistique Des Données Textuelles* (1128-1135). Universiteit Antwerpen.

Varios Autores. (2008). *Diccionario de términos clave de ELE*. SGEL; Instituto Cervantes.

Williams, M., & Burden, R. L. (1999). *Psicología para profesores de idiomas. Enfoque del constructivismo social*. Cambridge University Press.

SOLUCIONARIO

CAPÍTULO 1

1.5.1. ¿Qué ramas de la lingüística fueron pioneras en la recopilación de datos?

La antropología lingüística y la dialectología. Atención, fueron pioneras en la recolección de datos, pero todavía no hablamos de corpus.

1.5.2. ¿Cuáles son las ventajas del empirismo? ¿Y del racionalismo? ¿Y sus desventajas?

- Ventajas del racionalismo: creación de ejemplos para ejemplificar tus explicaciones; estos ejemplos son simples y limpios.
- Inconvenientes del racionalismo: posible interferencia del propio dialecto, idiolecto; verificación imposible; no sabemos si eso se dice en el mundo real, porque no lo hemos comprobado; dificultad de encontrar ejemplos “simples” en el mundo real: ¿cuántas veces hemos dicho *Juan come manzanas*?
- Ventajas del empirismo: datos auténticos; no hay interferencias de la propia persona o dialecto; disponibilidad de un mayor número de datos; es posible cuantificar los fenómenos (por ejemplo, un 20% de la gente usa la palabra *cuchufli*).
- Inconvenientes del empirismo: solo tenemos datos de lo que el corpus ha recogido; imposibilidad de tener evidencia negativa (esto tiene que ver con

el falsacionismo, propuesto por Popper); hay datos, pero no explicaciones; Abercrombie decía que no existe un ordenador capaz de analizar eso, pero ahora sí es posible.

CAPÍTULO 2

2.7.1. Fuera del mundo de la lingüística, la gente no suele pensar en datos o muestras de habla cuando decimos *corpus*. Vamos a comprobar esta intuición en un corpus de habla real, el *Corpus del Español* (CdE) (Davies, 2001). Para ello, vamos a usar un corpus en línea con interfaz de búsqueda; es decir, un buscador parecido a Google pero con textos ordenados. ¡Atención, el corpus de verdad son las muestras de habla que hay detrás, el buscador es solo una manera de consultar los datos de forma más fácil! El CdE es gratuito, pero después de un número de búsquedas pide que te registres, ¡no dudéis en hacerlo! El registro para uso de profesores o estudiantes es gratuito y lo vamos a usar mucho. Antes del CORPES, este era el único corpus con buscador lematizado para el español (veremos qué significa esto en el siguiente capítulo), eso hace que haya una infinidad de trabajos realizados con él.

1. En algún caso los datos se han podido actualizar, por lo que los resultados en años venideros pueden ser superiores a los que se dan aquí: 9837.
2. *Habeas corpus*, un término jurídico.
3. La fiesta religiosa del Corpus Christi, seguido de *habeas corpus* y en tercer lugar *corpus* en el sentido lingüístico, sobre todo con referencias a Google NGrams, un sistema de búsqueda que veremos más adelante en el curso.

2.7.2. Ve al corpus SOL (<https://spraakbanken.gu.se/korp/?mode=spanish#?lang=en&cqp=%5B%5D>) y busca las siguientes palabras en sus tres subcorpus: *presidente*, *presidenta*, *internet* e *email*.

1. *Presidente* aparece, *presidenta* también aunque con muchísimas menos ocurrencias. No aparecen ni *internet*, ni *email*, porque el SOL es un corpus que se recopiló antes de que esas palabras se usaran de manera generalizada.
2. El COR 92 (que es el *Corpus oral de referencia del español contemporáneo*). Es más reciente y da formas del oral.

CAPÍTULO 3

3.9.1. Ve a la web del CORPES.

1. Frecuencia absoluta: 17497216. Frecuencia normalizada: 62 207,38 casos por millón. Está normalizada por millón. Las palabras gramaticales son más frecuentes en la lengua (y por lo tanto también en cualquier corpus). Por ejemplo, siempre que aparece un artículo determinado será *el*, *ella*, *ellos* o *ellas*, solo hay cuatro posibilidades, mientras que para la aparición de un nombre hay miles.
2. *Amigas* nunca podrá ser un lema, es una forma del lema *amigo*. Así que la búsqueda como lema no devuelve resultados.
3. Por forma, ya que solo nos interesan las ocurrencias en femenino. La búsqueda por lema sería *juez* y aparecerían resultados con *juez*, *jueza*, *jueces* o *juezas*.

3.9.2. Ve al *Corpus Diacrónico del Español* (CORDE) y realiza una consulta de la palabra *cambera*.

1. 13 casos en cuatro documentos. No, no lo está. Los porcentajes que aparecen se refieren al 100 % de las apariciones de *cambera* por año, país y género, y no a la frecuencia de *cambera* en el total del corpus. Esta sería 13 apariciones de la palabra entre 250 millones de palabras que tiene el CORDE, 0,000000005 %.
2. Es típica de España. Además, a partir de la búsqueda en internet de la biografía de los autores, se puede ver que es típica de Cantabria y Palencia. Esto último queda ratificado al buscar la palabra en el DRAE, donde figura, como segunda acepción: 2. f. **rur. Cantb.** Camino de carros.
3. El CORDE (como el CREA) son corpus no lematizados. Permiten solo la búsqueda por forma.

3.9.3. Busca en el CORPES las colocaciones (en el CORPES las llaman *coapariciones*) de categoría VERBO de la palabra *posibilidad*.

1. *Barajar*, *sopesar*, *descartar*, *sondear*, *explorar*, *coartar*, *fantasear con* + la posibilidad.

3.9.4. Ahora vamos a buscar colocaciones en el *Corpus del Español* (web-dial).

1. *Cosas*, *decir*, *q* (abreviatura de *que*), *saber*, *comer*, *ver*, *resultados*, *dinero*, *sexo* y *vivir*.
2. Se pueden aprovechar los colocados más frecuentes para enseñar esos y no otros cuando se introduce el verbo *querer* en las clases.

3.9.5. Ve al formulario de búsqueda avanzada de ADESSE (<http://adesse.uvigo.es/data/avanzado.php>). Busca por rol semántico *sensación*. Al *experimentador*. Pulsa en REALIZAR BÚSQUEDA y, después, VERBOS. ¿Por qué podría ser interesante presentar esos verbos juntos en ELE?

El resultado de la búsqueda son verbos *experimentador* objeto y no sujeto, verbos del tipo *gustar*. Estos verbos tienen una construcción ajena a muchas lenguas. Pensemos en el inglés *I like cinema* (el yo *experimentador* es sujeto) que podríamos traducir por *#yo gusto del cine*, frente a la construcción del español *me gusta el cine* donde el sujeto gramatical es *cine*. Enseñar verbos como *interesar* o *importar* junto a otros como *gustar* o *encantar* que siguen el mismo patrón sintáctico puede ayudar a que los alumnos capten el matiz de objeto-experimentador y los aprendan más rápidamente.

CAPÍTULO 4

4.7.1. Busca la información de diseño del corpus CREA, disponible aquí: <https://www.rae.es/recursos/banco-de-datos/crea-escrito>. ¿Crees que es un corpus equilibrado respecto al origen de las obras? ¿Por qué?

No, porque el 50 % de los textos son de España, pero el número de hablantes de España representa el 9 % de la población. Así que un corpus representativo debería contener solo un 9 % de textos de España.

4.7.2. En el mundo del ELE, ¿un corpus de errores de alumnos de L1 de todas las lenguas del mundo estaría equilibrado? ¿Sería útil?

Sí, pero sería útil solo si estuviera codificado con la L1 del estudiante. Solo de esa manera, podríamos clasificar los errores según la L1 del estudiante, algo que es primordial en el aprendizaje de segundas lenguas.

4.7.3. El corpus CLUVI, recopilado por la Universidad de Vigo, es el corpus más completo en español de lenguajes especializados y traducción. Es uno de los corpus que se pueden usar en español con fines específicos. Para ver algunas de sus funcionalidades ve a la web: <http://sli.uvigo.es/CLUVI/index.php#>. En la parte superior izquierda, PESCUDAR TEXTOS, encontrarás el acceso a sus corpus organizados por tipo lengua especializada y lenguas:

1. *Sentencia o resolución.*
2. *Totally* se traduce por *totalmente* o *completamente*, también como, *todo* y, a

veces, con otras formas como *cuán*. *Completamente* aparece traducido en inglés como *all, altogether, totally* y *completely* (en menor medida). Esto nos muestra que, aunque hay palabras equivalentes, un cognado en la lengua con un significado muy parecido, su distribución de uso es diferente. Por lo tanto, en clase, incluso los cognados que no son falsos amigos pueden necesitar instrucción específica.

3. En texto literario el único caso es *ignorar*. En subtítulos: *saltar, ignorar, rechazar* u *omitir*.

4.7.4. Hace unos años, el corpus de libros publicados de Google Books se habilitó como corpus a través de una interfaz de búsqueda (<https://books.google.com/ngrams>). Como explicaban en la charla TED recomendada en el capítulo 2, este corpus es uno de los más visuales para ver cambios culturales en español. En el capítulo 8 veremos cómo llevarlo a clase de ELE, pero de momento, vamos a explorarlo. En la parte superior, escoge el corpus “Spanish 2012” y busca en esa interfaz de manera conjunta la palabra *médico* y *médica* y observa su frecuencia a través del tiempo.

1. Al uso de *médica* como adjetivo, como, por ejemplo, en *cita médica*.
2. La mayor parte de los casos de *médica* son como adjetivo. Resultado en la figura 8.

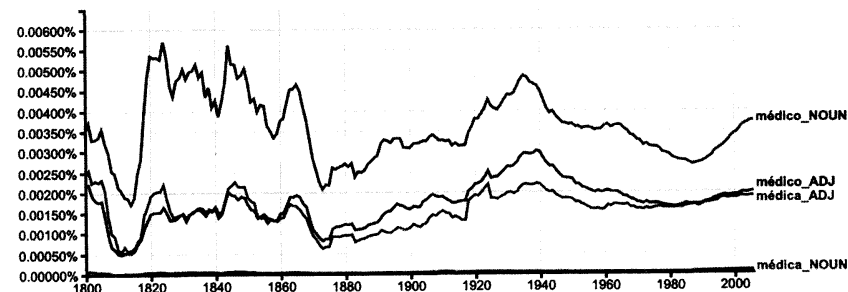


Figura 8. Resultado de la búsqueda en Google Ngrams (Spanish, 2012) de las diferentes opciones de *médico*

Como adjetivo, no hay diferencias notables entre el número de ocurrencias (frecuencia) del masculino y del femenino, puesto que el género es únicamente una cuestión de concordancia gramatical. En el caso de los nombres, el género hace también referencia a una realidad de sexo. En España, no hubo mujeres licenciadas

en Medicina hasta el siglo XIX y no hubo médicas de manera generalizada hasta los años 80 del siglo XX (ver la figura 9 para más detalles).

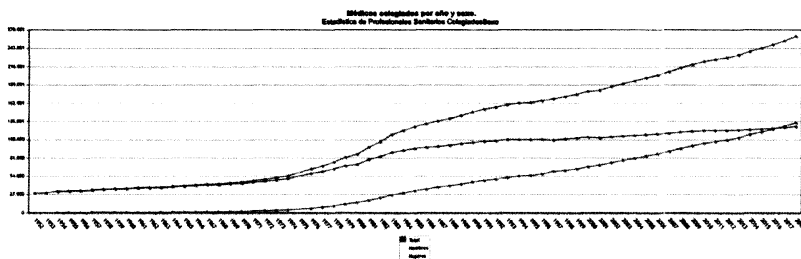


Figura 9. Número de colegiados en Medicina por sexo y año. Fuente: Instituto Nacional de Estadística (<https://www.ine.es>).

Incluso cuando ha habido médicas de manera generalizada, muchas de estas preferían el título de *médico* al de *médica*, igual que muchas poetisas prefieren que se refieran a ellas con este título al de *poetisa*. En 2016, se alcanzó la paridad en el número de colegiados de Medicina. Habrá que esperar para ver si esos resultados se reflejan en los corpus con paridad también en el término *médica* o las profesionales de la medicina siguen prefiriendo llamarse *médicos*. Los casos en los que se habla de *la médica* (sustantivo) de principios del siglo XX pueden ser en el sentido original etimológico (“mujer del médico”, recogido por el DRAE) o errores en la lematización del corpus. Si se hace una búsqueda de *médica* en el corpus de 2009, en la parte inferior, aparecen los extractos de los libros donde se ha encontrado la palabra (una especie de KWIC). Allí encontraremos que, en el año en el que el corpus tiene un pico de *médica* usado como nombre (1804), se encuentra lo siguiente: *materia Médica*, *Real Academia Médica*, *Anatomía Médica*. Es decir, se trata de casos de *médica* como adjetivo. ¿Por qué aparecen como nombre, entonces? Todos estos casos confunden al POS-tagger porque están escritos en mayúscula, y, normalmente, aparecen en mayúscula solo los nombres propios.

CAPÍTULO 6

6.5.1. Visita la página <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>. ¿Cuántos corpus tienen el español como lengua de estudio (*target*)? ¿Y cómo L1 (*first language*)? ¿Cuál de estas es más interesante para los profesores de ELE? ¿Por qué?

En 2019, como *target*: 8 multilingües, 12 monolingües. Como L1, 15 corpus. En ELE, el español es la lengua objetivo o meta, los estudiantes pueden tener cualquier L1.

6.5.2. Visita la página <https://slabank.talkbank.org/access/>. ¿Cuál de los corpus orales que hay disponibles te parece más interesante?

Respuesta abierta. Dentro de la oferta de corpus, Spanish o Spanish-Nebrija. Dependerá del interés de los estudiantes, por ejemplo, en ELE para adolescentes a un grupo de L1 concreto o en general.

6.5.3. Busca en el <http://cartago.llf.uam.es/exist/rest/db/corpus/errores.html> por TIPO DE ERROR “indicativo/subjuntivo” y LENGUA MATERNA “Francés”. Después de ver esos datos, ¿qué tema repasarías en clase?

El error que más veces aparece en el corpus tiene que ver con la formación del verbo *creer* + indicativo. Incidiría en que, en español, el verbo *creer* en afirmativo se construye con indicativo.

6.5.4. Busca ahora resultados de errores de concordancia de número en estudiantes cuya lengua materna es el inglés. ¿Qué errores son más comunes? ¿A qué rasgo lingüístico del inglés se pueden deber (esta pregunta presupone que tienes un nivel de inglés de al menos B1)? Un estudio completo y extenso sobre este tema podría dar lugar a un trabajo de curso que fuera: “La concordancia de número en la interlengua en español de aprendientes anglohablantes”.

Aparecen nombres incontables en inglés (*verduras* (por *vegetables*), *arroz*) y la palabra *gente* que en inglés se traduce como *people*, que es un nombre plural, mientras que en español *gente* concuerda en singular.

6.5.5. Ve al CEDEL (<http://cedel2.learnercorpora.com>), y entra o descarga cualquier producción. ¿Qué datos han incluido en el marcaje del corpus?

ID, Participant type (nativo o extranjero), Native Language (L1), Sex, Age, Age Started Spanish, Years Studying Spanish, Self Evaluation Spanish, Placement Test Percentage Score (0-100), Stay Spanish Country, Stay Length (in months), Other Foreign Language, Foreign languages, Self Evaluation Other Foreign, Composition theme, Word count (número de palabras escritas en el ejercicio).

CAPÍTULO 7

7.5.1. Estás preparando una unidad didáctica de A1 que tiene como tema “la familia” y has llegado al límite de piezas léxicas recomendadas para una sesión. Tienes que elegir entre las palabras *primo* y *sobrino*. ¿Cuál elegirías? Justifícalo por su frecuencia en corpus. Tendrás que elegir un corpus lo suficientemente grande, para que tus resultados sean representativos.

Teniendo en cuenta su frecuencia de uso, incluiría la palabra *primo*, ya que es más frecuente. En concreto, el lema *primo* tiene una frecuencia de 57,43 casos por millón en el CORPES, mientras que *sobrino* tiene una frecuencia de 21,27 casos por millón en el CORPES. En el corpus del español NOW, *primo* tiene una frecuencia de 66 518 ocurrencias y *sobrino*, de 59 182.

7.5.2. Busca en el CdE (NOW) la palabra *falda* y obtén la lista de concordancias.

1. En A1 se incluirían solo *larga* o *corta*. En un nivel B2 sería discutible dependiendo de el contenido del tema, pero se podrían incluir modificadores como *falda tubo* o colocaciones como *de rayas*, *de lunares*, que son altamente productivas porque se pueden aplicar a otras prendas de ropa.
2. El resultado devuelve: larga, negra, corta, plisada, blanca, escocesa, azul, amplia, roja, asimétrica, estampada, recta, gris, vaporosa, escolar, verde, vaquera, voluminosa, tradicional, abullonada. Lo habitual en un A1 sería incluir los colores y *corta* y *larga*. Sin embargo, se podría ampliar para aprovechar vocabulario que habrán visto en la misma unidad o conocen previamente y usar también *vaquera* o *escocesa*. Es llamativa la alta frecuencia de *plisada*, que, sin embargo, no se incluye en los programas de ELE.
3. En ese caso, incluiría los 20 pero creando una unidad didáctica en la que los alumnos los clasifiquen por “colores”, “tipos de falda por forma”, “materiales”, “estampados”...

7.5.3. Ve al corpus Valesco (<http://www.valesco.es/corpus/corpusrch.php>) y realiza una búsqueda por intervención con el contenido *es que*. ¿Cuál es la función de las ocurrencias de *es que* en el corpus?

La consulta se debe realizar por intervención y no por forma o lema, ya que estamos buscando una secuencia y no una sola palabra. Si se quiere repetir la

búsqueda, hay que pulsar antes RESETEAR, de lo contrario la interfaz no responde.

La función de *es que* es de presentar argumentos. En ELE se suele incluir en los programas de estudios como introductor de justificaciones, por ejemplo: —¿Vienes la cine?, —Es que tengo que estudiar. La construcción tiene además una entonación prototípica asociada: el tono sube ligeramente al final, por lo que tener un corpus oral para escucharlo es ideal. El corpus permite escuchar la intervención tras pulsar en la lupa que abre la intervención en una nueva página con toda su información asociada.

Corpus Val.Es.Co. 2.1

The screenshot shows the Valesco search interface. At the top, it says 'Corpus Val.Es.Co. 2.1'. Below that, there's a search bar with the text 'intervenciones consulta' and a 'Buscar' button. To the right of the search bar, there's a link that says '# / intervenciones consulta / Buscar'. Below the search bar, there are several filters: 'duracion' with a range from 0 to 10, 'orden' with a dropdown menu, 'intervencion contains' with a text input field containing 'es que', 'sexo' with radio buttons for 'hombre' and 'mujer', 'nivel' with radio buttons for 'alto', 'medio', and 'bajo', and 'edad' with radio buttons for '18-25', '26-55', and 'mayor de 55'. At the bottom, there are two buttons: 'Buscar' and 'Resetear'.

Figura 10. Interfaz de Valesco con los parámetros de búsqueda necesarios

7.5.4. Tus alumnos te han pedido una clase monográfica de comida. Elige un ingrediente o un plato típico (por ejemplo, *tortilla* o *huevo*) y extrae de un corpus sus colocaciones/coapariciones para obtener el contenido léxico de la clase. Si usas el CORPES, no olvides ordenar por frecuencia para poder ver las coapariciones más habituales. Tienes dos enfoques posibles para la clase: 1) expresiones fraseológicas (*darle la vuelta a la tortilla*, *costar un huevo*); 2) colocaciones (*tortilla de maíz*, *huevo pochado*).

Huevo: huevo duro, huevo cocido, huevo frito, huevo batido, huevo de Pascua, huevo kinder, clara de huevo, cáscara de huevo, huevo entero, huevo gigante, huevo de gallina...

7.5.5. Busca en el CdE la palabra *cuando* seguida de verbos en imperfecto, elige una (la que te parezca más útil para llevar a clase)

de las tres colocaciones resultantes de más frecuencia y mira las 10 primeras líneas de concordancias. Adapta tres para que los ejemplos obtenidos sean aptos para llevar a clase.

Las tres entradas con más frecuencia son *cuando era*, *cuando estaba* y *cuando tenía*. Las tres son colocaciones que aparecen fácilmente en un A2, cuando se introducen las expresiones temporales y, específicamente, la construcción *cuando* + imperfecto + imperfecto. Por ejemplo, *cuando era pequeño* o *quería ir a la luna*.

Estas son las 10 líneas de concordancias de *cuando era* en el momento de la búsqueda⁴. Hay que tener en cuenta que los corpus se van actualizando por lo que es posible que vuestra búsqueda dé un resultado diferente.

1	dentro de el PSC, y que milita en el partido que contribuyó a fundar [REDACTED] joven, en plena clandestinidad: "Es un partido que hace de la
2	, lo que es muy grave, porque él dice que supo en la época [REDACTED] militar, entonces él reconoce la relación con los calabozos de la dictadura.
3	« es echar un cable a los compañeros. Porque me recuerda (Pedri) [REDACTED] más joven. Me veo reflejado en ello y mi misión con estos compañeros
4	dentro de el PSC, y que milita en el partido que contribuyó a fundar [REDACTED] joven, en plena clandestinidad: "Es un partido que hace de la
5	cadena también ha declarado sobre el fallecimiento de el actor en un comunicado: # [REDACTED] joven, fue impulsado por un fuerte deseo de hacer una diferencia en la
6	de sumario. # Las fuentes consultadas explican que el kilo de cocaína fue intervenido [REDACTED] transportado en un vehículo desde la Costa de el Sol hasta Linares. Los
7	a 90 años de edad ha sido trasladada a un centro hospitalario tras sufrir heridas [REDACTED] evacuada por su hijo de la vivienda en la que ambos residían en la
8	memoria. # Lefebvre, el reaccionario francés a quien nuestro Papa Benedicto PP XVI [REDACTED] Prefecto de la CDF tuvo la deferencia de animar lo a que presentara sus
9	ir hacia el individualismo y el egocentrismo. Los ordenadores lo hacen todo fácil. [REDACTED] niño, mis referencias eran mis padres, el sacerdote y la escuela.
10	puestos de escalafón? # –Bueno, he conseguido mucho más de lo que imaginé [REDACTED] pequeño. Yo soñaba con ser torero y tomar la alternativa, pero nunca

Un ejemplo de adaptación apto para clase sería:

1. *Cuando era niño, mis referencias eran mis padres.* (Basado en la línea 9, elimino *sacerdote* y *escuela* por ser un ejemplo menos típico en la actualidad).

⁴ Una línea de concordancia puede aparecer más de una vez en la lista de resultados (como aquí en 1 y 4) cuando provienen de dos fuentes diferentes, por ejemplo, dos ediciones diferentes del mismo libro o dos páginas web con el mismo contenido.

2. *Cuando era niña, soñaba con ser astronauta.* (Basado en 10, cambio el género, porque no hay ningún ejemplo en femenino y *torero* por *astronauta*, para no proporcionar únicamente ejemplos de tópicos).
3. *Cuando era joven, fundó un partido.* (Basado en 1. Atención, en este caso se incluye el indefinido por lo que estamos hablando de la construcción *cuando* + imperfecto + indefinido).

CAPÍTULO 8

8.8.1. Piensa un conjunto de búsquedas que podrían dar juego en un tema dedicado a las profesiones para mostrar la caída y auge de algunas de ellas. Por ejemplo, si buscáramos *pregonero*, veríamos que su uso ha caído. También lo puedes enfocar al género, a partir de búsquedas como *secretario* y *secretaria* (si buscáis esta en concreto, os llevaréis una sorpresa y veréis que no tiene sentido llevarla a clase).

Ejemplos de empleos de nueva creación y, hoy en día, muy habituales serían *informático* y *diseñador*. Se puede encontrar con la búsqueda lematizada "informático_NOUN" o buscar de manera conjunta y separada por género "diseñador_NOUN, diseñadora, diseñador". La búsqueda cajera también dará una muestra de esto mismo, si se busca "cajero_NOUN" nos encontraremos que muchas ocurrencias no se refieren a la profesión así, que aparecen ya en el siglo XIX.

Para buscar profesiones y su relación con el género, la búsqueda será "secretario, secretaria". Lo curioso de esta búsqueda es que contrariamente a lo que nos diría nuestra intuición, los secretarios son muy habituales. Esto se debe a colocaciones como *secretario de estado*, que son ocupadas mayormente por hombres y salen muy habitualmente en libros y periódicos. Si lo que queremos mostrar en clase es que la palabra *secretario*, como *azafato*, es poco habitual y que se prefieren formas como *asistente de vuelo* o *administrativo*, esta búsqueda no nos serviría para ilustrarlo, porque necesita que expliquemos demasiado. Nuestros ejemplos de clase tienen que ser inequívocos, mucho más de lo que es la vida real. Para ejemplos, equívocos tenemos las muestras de habla real, el trabajo de competencias, pero no la teoría.

8.8.2. En español existe una construcción de negación que permite combinar el adverbio con la presencia de otro adverbio o indefinido de sentido negativo, palabras como *nunca*, *jamás*, *tampoco*, *nadie*, *nada* o *ninguno*. Diseña una actividad de corpus para que los alumnos puedan deducir cuándo una de estas palabras negativas aparece acompañada

del adverbio *no*. Es probable que la búsqueda en corpus que les propongas contenga, además de la forma de la palabra de interés, el adverbio *no* como colocado.

La búsqueda se puede realizar en cualquier corpus lematizado y anotado, como el CORPES o el CdE. Habría que realizar dos búsquedas: una para el adverbio, por ejemplo, *nunca* y otra con el adverbio acompañado de *no* como colocado a bastante distancia, por ejemplo, a seis. El resultado razonado de la búsqueda es que el adverbio *no* aparece cuando la palabra negativa está pospuesta: *Nunca voy al cine* vs. *No voy al cine nunca*.

8.8.3. Piensa y anota la búsqueda que tendrían que llevar a cabo tus alumnos en el Corpus del español para obtener las colocaciones más habituales de *volverse* + adjetivo.

La búsqueda sería la siguiente: en la pestaña de colocados, “*volverse_V**” con un colocado a la derecha y ese colocado puede ser de cualquier tipo (*) y ninguno a la izquierda. En el momento de la publicación, en el subcorpus NOW, una colocación muy frecuente es *Volverse palestina*, es el título de una novela. Una vez descartado esto, las colocaciones más habituales son *volverse loco*, *rico*, *famoso*, *millonario*, *humano*. Otras colocaciones débiles pero que puede interesar a nuestros alumnos son *vegano*, *global*, *previsible*...

Word/phrase	POS	Frequency
loco	[*]	4
rico	[*]	3
famoso	[*]	2
millonario	[*]	1
humano	[*]	0
vegano	[*]	0
global	[*]	1
previsible	[*]	2
...	[*]	3
...	[*]	4

Figura 11. Modelo búsqueda en la interfaz.

En el CORPES, la búsqueda sería con lema *volver*, ya que *volverse* no está lematizado de manera diferente, y no permite acotar la categoría del colocado por lo que las primeras páginas de resultados obtenidos son palabras gramaticales e incluso signos de puntuación como “; a el y se de que”.

8.8.4. Propón una actividad en la que los estudiantes tienen que deducir el significado de una palabra a partir de las líneas de concordancias.

La respuesta a esta actividad es abierta, un modelo posible sería el que se da abajo.

- En casa, busca en el corpus CORPES (<http://web.frl.es/CORPES/view/inicioExterno.view>) la palabra *lechuga*: busca *lechuga* en lema y clica CONCORDANCIAS. Observa los contextos en los que aparece. ¿Qué crees que es?
- En clase, compara tu respuesta con la de tu compañero. ¿Creéis que significa lo mismo?

Otra opción menos exigente podría ser preguntar algo como: *Cómo*: ¿crees que es algo positivo o negativo?

8.8.5. Entra en el Atlas interactivo de la entonación del español y busca la entonación de las preguntas absolutas > de tipo neutro > preguntas informativas de Madrid (España) y Santiago de los Caballeros (República Dominicana). El tono está marcado por la curva negra horizontal ¿cuál te parece la diferencia más sobresaliente entre las dos? Madrid está en el mapa de España (http://prosodia.upf.edu/atlasentonacion/mapa_e.html), Santiago de los Caballeros en el mapa de América (http://prosodia.upf.edu/atlasentonacion/mapa_l.html).

Las curvas se pueden consultar en http://prosodia.upf.edu/atlasentonacion/enquestes/espanol/santiago_caballeros/index.html y <http://prosodia.upf.edu/atlasentonacion/enquestes/espanol/madrid/index.html>. La mayor diferencia es que la curva de Madrid tiene una subida al final, mientras que la curva dominicana se mantiene alta y desciende al final.

CAPÍTULO 9

9.5.1. Imagina que un alumno ha cometido los siguientes errores. Indícale las búsquedas que debería realizar y en qué corpus (CdE o usar la web como corpus mediante Google) para encontrar las formas correctas. Por ejemplo, **esquina del ojo* > CdE, busca “[N*] del ojo”. La referencia al final de los ejemplos indica el corpus de dónde se han extraído.

- Google, búsqueda entrecomillada “La noche doy una vuelta”.
- CdE (NOW o Genre), [V*] espectáculos. (Nota: aunque la producción no sea incorrecta, por algún extraño motivo solemos *asistir* a actos y, en cambio, *ver* espectáculos).

- Google, escribe “persona muy importante” y observa las opciones de autocompletado.
- CdE, [V*] siesta (Nota: la construcción con *hacer* se usa en Centroamérica, pero entonces se tiene que utilizar sin el determinante, es decir, o *dormir la siesta* o *hacer siesta*. El alumno es tunecino, por lo que es probable que todo su *input* sea de español de España. La colocación *hacer la siesta* es de uso común en español de Cataluña por interferencia del catalán).

CAPÍTULO 10

10.2.1. Crea una actividad en la que tus alumnos tengan que usar recursos de corpus para hacer una redacción. Pon especial cuidado en poner límites para que no busquen cada una de las palabras y colocaciones de la redacción.

Ejercicio de respuesta abierta. Un modelo sería el siguiente:

Escribe la biografía de tu personaje histórico favorito. Busca en un corpus los verbos con preposición usados para comprobar que son correctos.

CAPÍTULO 11

11.5.1. Plantea una secuencia didáctica para tratar el uso y forma del subjuntivo. No es necesario que la desarrolles en su totalidad, solo apunta los temas que tratarías y qué corpus usarías y por qué.

Respuesta abierta, se puede tomar cómo modelo la secuencia didáctica que se ofrece en el tema.

El corpus usado tiene que incluir anotación gramatical que permita buscar verbos en subjuntivo, el CdE serviría perfectamente.

GLOSARIO

ADESSE: Corpus en el que se pueden consultar construcciones con anotación semántica, por ejemplo, aquellos verbos que tienen un sujeto experimentador o verbos que se construyen con complemento directo e indirecto.

Anotación: Información gramatical (fonética, morfosintáctica, semántica...) incluida en un corpus.

Aprendizaje basado en datos (*data-driven*): Acercamiento al aprendizaje deductivo, que consiste en que el alumno infiera las reglas a partir de la observación de muestras de habla.

Aprendizaje inductivo: Método de aprendizaje en el que el alumno observa y analiza un fenómeno lingüístico para llegar a deducir la regla gramatical que subyace.

Aprendizaje por descubrimiento: Método de aprendizaje en el que el estudiante debe adquirir los conocimientos por sí mismo a partir del *input*.

BYU: Colección de corpus del inglés, cuya interfaz de búsqueda usa también el Corpus del Español.

Clase invertida: Tendencia de la pedagogía en la que se propone que los estudiantes preparen el tema en casa antes de la clase para poder dedicar el tiempo de clase a la discusión y profundizar en los contenidos .

COBUILD: Diccionario del inglés creado a partir del corpus del Bank of English.

Cognados: Palabras que comparten significado, ortografía y pronunciación similares en dos idiomas.

Colocación: Grupo de palabras que se combinan juntas de manera frecuente.

Concordancia: Contexto en el que aparece la palabra visualizado alrededor de esta.

CORDE: Corpus diacrónico de la Real Academia Española.

CORLEC: Corpus oral en un registro estándar.

CORPES: Corpus sincrónico consultable en línea de la Real Academia Española.

Corpus audiovisual: Corpus cuyas muestras de habla son en formato vídeo.

Corpus canónico: Corpus que contiene toda la obra literaria de un autor.

Corpus comparable: Corpus que contiene textos parecidos en dos o más lenguas, por ejemplo, anuncios de venta de pisos en español y en inglés, pero no son el mismo anuncio traducido.

Corpus crudo: Corpus que carece de anotación o marcaje.

Corpus de errores o de aprendientes: Corpus que recoge muestras de habla (escritas u orales) de estudiantes de español lengua extranjera.

Corpus de lengua especializada: Corpus que contiene muestras de un campo concreto, por ejemplo, lenguaje judicial o médico.

Corpus del Español (CdE): Primer corpus del español lematizado con interfaz en línea. Es el mayor corpus del español disponibles y el más usado en los estudios de lingüística.

Corpus diacrónico: Corpus que recoge muestras de un periodo de tiempo con el objetivo de ver la evolución de la lengua.

Corpus equilibrado: Corpus que cumple los criterios de diseño necesarios para ser representativo de una variedad de habla.

Corpus escrito: Corpus de muestras de habla cuyo soporte original es escrito, en línea o manuscrito.

Corpus general: Corpus cuya intención es la de representar la lengua general, no especializado.

Corpus monitor: Corpus que se actualiza periódicamente para mostrar la lengua lo más actualizada posible.

Corpus multilingüe: Corpus que contiene datos de más de una lengua.

Corpus oral: Corpus con muestras de la muestra oral, puede contener grabaciones y sus transcripciones o solo las transcripciones.

Corpus paralelo: Corpus que contiene el mismo texto en su versión original y al menos una traducción a otra lengua.

Corpus sincrónico: Corpus que contiene muestras de un solo periodo de tiempo. Ese periodo de tiempo puede ser actual o no. Es decir, también es sincrónico un corpus con muestras de los años 70.

CREA: Primer corpus sincrónico consultable en línea de la Real Academia Española (sin lematizar).

Dato lingüístico: Muestras de habla real, habladas o escritas.

Derivación: Proceso de creación de palabras mediante la adición de un morfema a la raíz.

Empirismo: Creación de conocimiento a partir de los datos.

Entonación circunfleja: Entonación descendente.

Enfoque comunicativo: Método de enseñanza de lenguas cuyo objetivo es la comunicación efectiva fuera del aula.

Español para fines específicos: Subcampo del ELE que se centra en la enseñanza del español para ejercer una profesión en concreto. Son especialmente relevantes los campos del turismo, los negocios o las relaciones internacionales.

Familia léxica: Conjunto de palabras que se forman sobre la misma raíz.

Forma: Palabra flexionada.

Frecuencia normalizada: Número de veces que una palabra (*token*) aparece en un texto dividido entre un número concreto o el tamaño total del corpus.

Frecuencia: número de veces que una palabra (*token*) aparece en un texto.

Google NGram Viewer: Buscador en los libros digitalizados por Google Books, probablemente la interfaz de búsqueda más intuitiva de los corpus en español.

Hápax: Palabra que se ha documentado solo una vez en un corpus. Es decir, que tiene una frecuencia no normalizada de 1.

Idiolecto: Variedad lingüística propia y personal de cada individuo.

Interlengua: Sistema (lengua) que se usa durante el aprendizaje de una lengua extranjera y que está caracterizado por la interferencia de la L1 y por la simplificación del sistema.

Jerga: Variedad lingüística específica de algunas profesiones y grupos sociales.

Lema: Forma no marcada de una palabra, palabra que encabeza el diccionario.

Lingüística de corpus: La metodología o el marco teórico de la lingüística que hace uso de corpus.

LOB: Corpus histórico del inglés británico recopilado en los años 70 y que contiene un millón de palabras.

Marcaje o codificación: Información externa de las muestras de un corpus donde se indica la procedencia de la muestra de habla, su autor y características, su canal original, su época, etc.

Método científico: Metodología para verificar hipótesis basada en la observación, estudio de los datos y refutación de las teorías.

Método gramática-traducción: Método de enseñanza de lenguas a partir de la traducción de textos escritos.

Ocurrencia: Cada una de las apariciones de una forma o lema en un corpus.

Racionalismo: Creación de conocimiento a partir de la razón o la intuición.

Registro: Variedad de lengua adecuada para un contexto social concreto. Por ejemplo, en televisión se habla un registro estándar (formal), mientras que en el ámbito familiar se usa un registro informal.

Sol: Uno de los primeros corpus disponibles en línea para el español. Contiene principalmente datos de novelas. Hoy en día ha quedado algo desactualizado.

Terminología: Conjunto del léxico usado por una disciplina.

Unidad fraseológica: Combinación de palabras idiomática que aparece en un determinado orden y cuyo significado puede no ser deducible de sus partes, por ejemplo, *pie de rey*.