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 ORTOTIPOGRÁFICA: Marina López

DISEÑO DE CUBIERTA E INTERIORES: Laurianne López Barrera

MAQUETACIÓN: Aleix Tormo

ILUSTRACIÓN: Laurianne López Barrera

978-84-18625-32-9
Impreso en la UE

difusión
Centro de Investigación y Publicaciones de Idiomas, S. L.

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 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.

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*, *enfoque 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 Valdehíta (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

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 *cuchuflú*, 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).

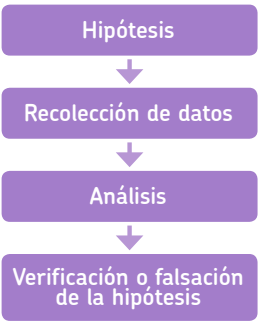
Esto crea una gran separación en lingüística que tiene continuidad hasta nuestros días: lingüistas racionalistas, que usan su intelecto para analizar; y lingüistas empiristas, que se basan en los datos.

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 Linguee 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?