El *Corpus de Aprendices de Español* (CAES) y sus aplicaciones para la enseñanza/aprendizaje del español como lengua extranjera

The *Corpus de Aprendices de Español* (CAES) and its applications to the teaching/learning of Spanish as a foreign language

Ignacio Palacios Martínez (USC)
Francisco Mario Barcala (NLPgo)
Guillermo Rojo (USC)

RESUMEN:

El objetivo de este trabajo es presentar las características principales del CAES y sus aplicaciones en diversos ámbitos de la enseñanza y aprendizaje del español como lengua extranjera. En la primera parte se explica el origen de este proyecto así como las propiedades más importantes del corpus: diseño y estructura, proceso seguido para su compilación, tamaño, niveles y lenguas origen representadas. Se incluye aquí también una breve información referida a los procesos de etiquetación morfosintáctica y lematización, así como a las funcionalidades de la herramienta de consulta.

La segunda parte se centra en su totalidad en las aplicaciones y posibilidades de explotación del corpus que van desde investigaciones sobre las dificultades que tienen los aprendices en su aprendizaje y estudios contrastivos de interlengua hasta la elaboración de material de aula con datos extraídos del propio corpus y sus implicaciones para ámbitos como el diseño y desarrollo curricular, la formación del profesorado y la evaluación.

La aproximación adoptada es muy práctica de modo que las explicaciones teóricas van acompañadas de ejemplos ilustrativos que facilitan su comprensión.

Palabras clave: lingüística de corpus, corpus de aprendices, interlengua, español lengua extranjera.

ABSTRACT:

The objective of this paper is to present the main characteristics of CAES and its application to various areas of the teaching and learning of Spanish as a foreign language. The first section describes the origins of the project and its most salient features: design and structure, compilation process, size, learner-levels, and source languages represented. Also included here is some brief information on the processes of morphosyntactic tagging and lemmatisation, as well as a description of the functionalities of the query tool.

The second section focuses on its applications, including research into the difficulties that learners face during the language-learning process, contrastive studies of interlanguage, and the pro-

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.

duction of classroom material with data extracted from the corpus, as well as the implications here for areas such as curriculum design and development, teacher training, and evaluation.
The approach adopted is very practical, and theoretical explanations are accompanied by illustrative examples that allow for easy understanding by non-specialists.
Keywords: corpus linguistics, learner corpus, interlanguage, Spanish as a foreign language.

1. INTRODUCCIÓN

El propósito de este trabajo reside en describir las características principales del *Corpus de aprendices de español como lengua extranjera* (CAES) (<http://galvan.usc.es/caes>) prestando atención especial a su explotación y aplicaciones por parte de docentes, investigadores y especialistas en el diseño y elaboración de materiales para la enseñanza y el aprendizaje del español como lengua extranjera. La aproximación utilizada es eminentemente práctica de modo que contribuya a despertar el interés por su consulta y anime a su utilización como herramienta de trabajo tanto para el estudio, la clase o la investigación.

Este capítulo está organizado en dos secciones o apartados principales; así, mientras que en la primera parte se explican el origen y desarrollo de este proyecto con sus características básicas más relevantes, la segunda se centra en las posibles aplicaciones de este corpus para la investigación del proceso de aprendizaje del español como lengua extranjera, la elaboración de actividades de aula, el diseño y desarrollo curricular, la formación del profesorado y la propia evaluación del trabajo del alumno. En lo posible, se ha intentado proporcionar ejemplos prácticos a modo de ilustraciones de todas estas aplicaciones.

2. BREVE DESCRIPCIÓN DEL CAES

El *Corpus de aprendices de español como lengua extranjera* (CAES) es resultado del empeño personal de Francisco Moreno Fernández. En 2010, cuando fungía como director académico del Instituto Cervantes, nos encargó a los autores que preparásemos el diseño de un corpus de aprendices de español como lengua extranjera y de todos los procesos necesarios para desarrollarlo desde la recogida de muestras hasta su publicación. Presentamos el proyecto pocas semanas después y, tras los trámites administrativos oportunos, el Instituto Cervantes (IC) contrató con la Universidade de Santiago de Compostela (USC) el diseño, construcción, tratamiento lingüístico y explotación del CAES.

Analizadas las necesidades del IC para este proyecto, propusimos centrar el trabajo en estudiantes con seis L1 diferentes (árabe, chino mandarín, francés, inglés, portugués y ruso). Consideramos que, dada la implicación del IC en el proyecto, lo más adecuado era estructurar la recogida de pruebas en función de los niveles de conocimiento ya adquiridos por los participantes, con lo que quedaba claro desde el principio que no íbamos a tener estudiantes del

nivel C2, de muy difícil o imposible localización. Por fin, decidimos montar el proceso de recogida mediante una aplicación informática que, en un entorno controlado (idealmente un centro de recursos del IC), pidiera a los participantes que cubrieran un formulario con sus datos y luego escribieran, en la misma sesión de trabajo, las pruebas correspondientes a su nivel de conocimientos. Al final de la sesión, la persona encargada del control de las pruebas las remitía a un servidor de la USC añadiendo un parte de incidencias producidas.

El procedimiento utilizado tiene dos grandes ventajas. La primera de ellas consiste en que no es necesario realizar la ardua y costosa tarea de escanear, transcribir y procesar textos manuscritos. La segunda procede del hecho de que el trabajo se ha simplificado también al ser posible utilizar el formato digital más adecuado en cada fase del procesamiento, como veremos más adelante. Con ello, la carga de trabajo de esta parte del proyecto resultó mucho más llevadera. Las desventajas radicaban en la necesidad de que las pruebas se realizaran en una sala con, por lo menos, conexión a Internet en la que los participantes pudieran trabajar en computadoras del centro o bien con sus propias máquinas. Aunque las ventajas superan con mucho a los inconvenientes, es claro que estos requisitos pueden suponer una dificultad adicional, puesto que no todos los centros del IC ni de las universidades colaboradoras disponen de instalaciones con estas características que puedan usar con facilidad.

Las pruebas solicitadas consistieron en la redacción de tres textos de entre 30 y 200 palabras cada uno en el caso de los estudiantes de los niveles (ya logrados, como hemos dicho) A1, A2 y B1 mientras que ese número se reducía a dos pero con una extensión aproximada de 275-500 palabras para los alumnos de los niveles B2 y C1. Consideramos adecuado fijar los temas de esas composiciones, así que se les pedía reservar una habitación en un hotel, hacer una reclamación de equipaje a una compañía aérea, contar una película, etc. Es decir, temas centrados en aspectos corrientes de la vida y planteables, por tanto, con distintos grados de conocimiento lingüístico. Muchos de estos temas suponen, como es lógico, una concentración de los elementos léxicos en ciertas zonas, lo cual provoca dificultades de importancia si se pretende proyectar los lemas obtenidos en el análisis de las producciones a cuestiones relacionadas con el análisis general del vocabulario adquirido por los estudiantes.

El diseño del proyecto incluía, naturalmente, la anotación morfosintáctica y la lematización de las producciones. Tras el análisis detenido de un conjunto de pruebas correspondientes a diferentes niveles y L1, llegamos a la conclusión de que los resultados del análisis automático iban a ser muy deficientes aunque dedicáramos tiempo y esfuerzo a la construcción de un corpus de entrenamiento, que debería tener muestras amplias por la multiplicidad de niveles y lenguas de origen. Optamos, en consecuencia, por utilizar una aplicación de anotación ya construida y concentrar el esfuerzo en la corrección manual (desambiguación) de sus resultados. Freeling, de uso libre, nos pareció la mejor opción y, tras algunas pruebas iniciales, fue la que

utilizamos finalmente, con un par de adiciones. Por una parte, Susana Sotelo de la USC preparó una rutina que combinaba la segmentación de oraciones con su *tokenización* y análisis con Freeling. Para mayor comodidad en la desambiguación, permitimos que Freeling seleccionara la etiqueta que presentaba los mejores resultados, pero añadiendo luego todas las etiquetas posibles para cada uno de los elementos resultantes del análisis automático. Por otro lado, María Paula Santalla del Río, también de la USC, construyó otra rutina para reconvertir algunas de las etiquetas usadas por la versión general de FreeLing a otras más conformes con el sistema de anotación usado habitualmente en nuestro grupo de la USC.

El paso siguiente consistió en el desarrollo de una aplicación que permitiera corregir los resultados automáticos haciendo las sustituciones, adiciones y eliminaciones tanto de etiquetas como de lemas. La tarea de corrección, dirigida por Paula Santalla, fue realizada por Alba Fernández Sanmartín y Marlén González González González, lingüistas contratadas directamente en el proyecto. Se hizo una fase de análisis por separado de los mismos textos, para comprobar el etiquetario, resolver dudas, fijar criterios y tomar decisiones en los casos dudosos.

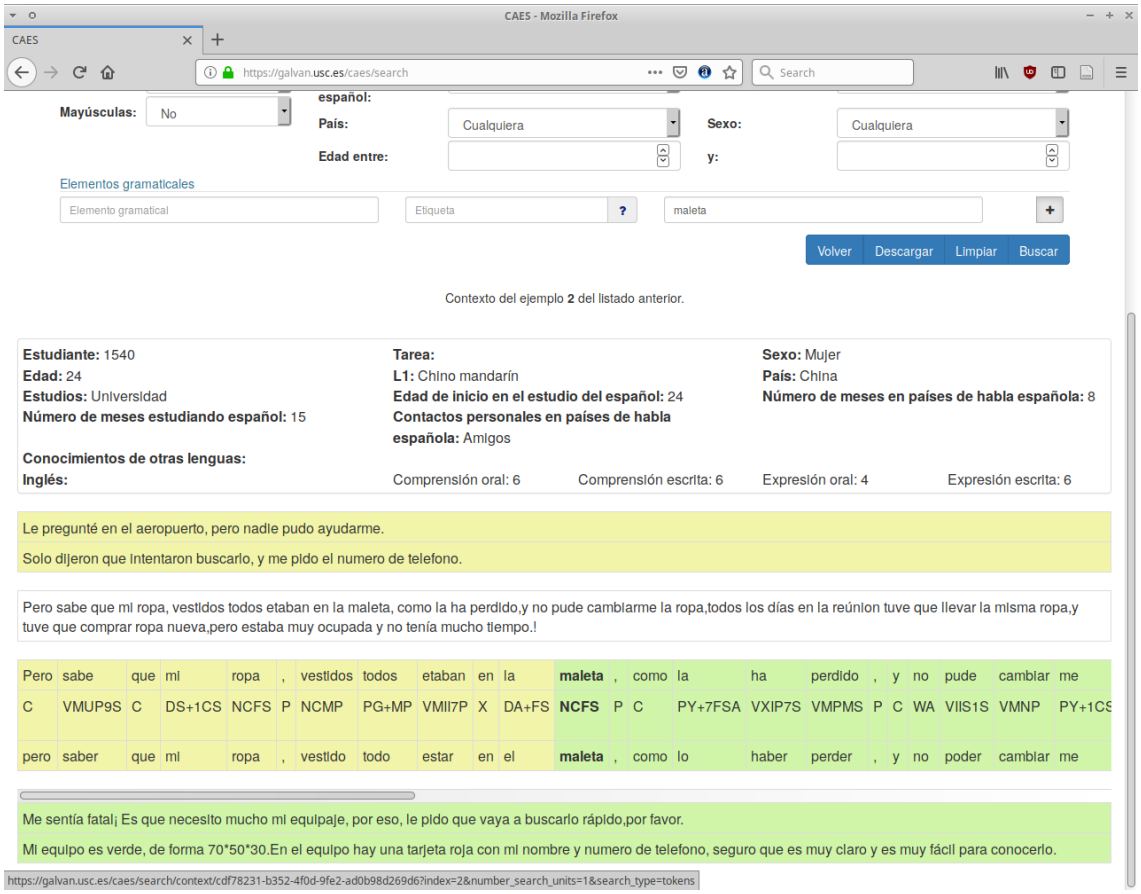
El resultado de todo ello es un corpus de unos 600 000 elementos lingüísticos etiquetados y lematizados por lingüistas expertos en estas tareas. Se atribuyó etiqueta y lema también a aquellas formas inexistentes en español, de modo que *trayó*, por ejemplo, es etiquetada como la tercera persona de singular del pretérito de indicativo del verbo *traer*.

El paso final fue la construcción de una aplicación de consulta, desarrollada por Francisco Mario Barcala (NPLgo), que permitiera trabajar con todos los rasgos de los aprendices incluidos en la configuración del corpus (L1, nivel, sexo, edad, etc.) y añadir toda la potencia derivada de la anotación morfosintáctica y la lematización. Es posible, por ejemplo, recuperar todos los casos del verbo *ir* seguido del infinitivo de cualquier verbo a una distancia de uno o dos elementos (para que devuelva tanto casos del tipo *voy a salir* como *voy salir*). Como se ha indicado ya, esa búsqueda abstracta, basada en la consideración unitaria de todas las formas del paradigma del verbo *ir* y la detección de la forma de infinitivo de cualquier verbo, puede referirse a todos los estudiantes de nivel A1, a todos los estudiantes con portugués como L1 o a los estudiantes con nivel A1 y portugués como L1, que es el tipo de opción que proporciona habitualmente los datos necesarios para la investigación en este terreno. En los apartados 3.1 y 3.2 se muestran los resultados obtenidos en el análisis de algunos fenómenos como el mencionado.

Además de proporcionar la estadística simple, la ampliada y las concordancias, la aplicación de consulta permite acceder a toda la información correspondiente a la oración en la que se encuentra un determinado elemento. Como muestra la pantalla reproducida en la figura 1 se obtiene la secuencia de formas escritas originalmente, los elementos lingüísticos que la

constituyen y los lemas a que pertenecen, con lo que se tiene acceso a la totalidad de los datos manejados, tanto en su forma original como con las capas de información añadida. Finalmente, todo ello es descargable en formato tsv (es decir, formato de texto con campos separados por tabuladores), que permite tanto su manejo directo como su integración en una hoja de cálculo o una base de datos.

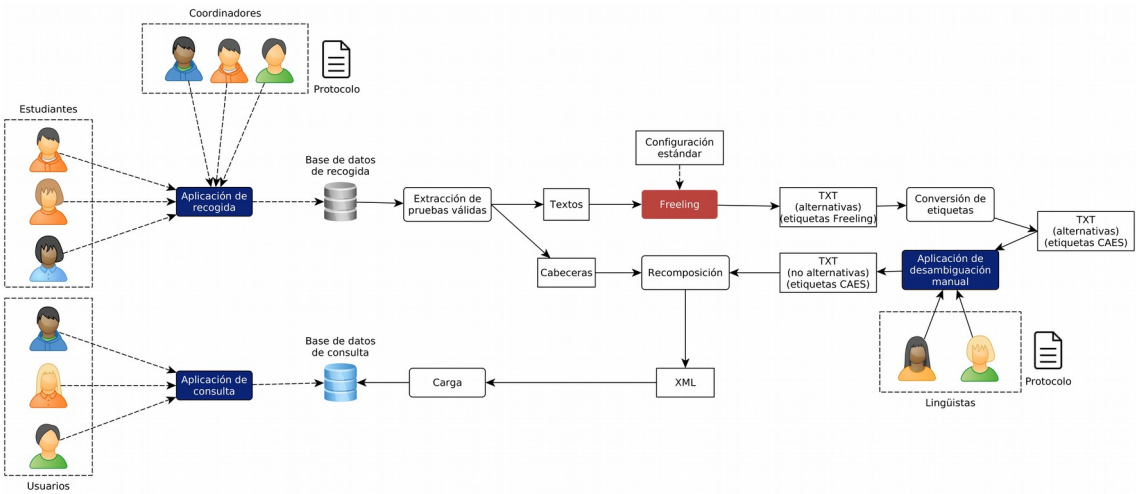
Figura 1: Ejemplo de los resultados obtenidos de una consulta simple en el corpus



En la figura 2 se muestra el flujo de procesamiento que han seguido las pruebas, desde que se organizaron las sesiones de recogida hasta que estas pasaron a formar parte de la aplicación de búsqueda. Como primer paso, los coordinadores habilitaban en la aplicación *online* de recogida diferentes sesiones para la realización de pruebas y los estudiantes realizaban las pruebas correspondientes que, al final de cada sesión, eran almacenadas en una base de datos relacional. Una vez eliminadas las muestras inválidas (duplicadas, incompletas o incluso vacías, entre otras causas), los textos sin las cabeceras fueron etiquetados con Freeling y las etiquetas asignadas fueron convertidas en las equivalentes que se utilizaron en el proyecto, como ya hemos descrito anteriormente. A continuación, utilizando una herramienta desarrollada específicamente para esta tarea, se desambiguó manualmente el resultado de la etiquetación automática y, finalmente, las pruebas, ya etiquetadas y desambiguadas, se volvieron a unir a sus

metadatos para generar un documento XML, lo que facilitó la realización de la carga de las pruebas en una nueva base de datos específicamente diseñada para permitir realizar las búsquedas de la aplicación de consulta.

Figura 2: Flujo de procesamiento seguido por las pruebas desde el inicio hasta el final del proceso



3. APLICACIONES DEL CAES

Si bien existe un gran número de publicaciones en torno a las aplicaciones de la Lingüística de corpus y de los corpus de carácter general, por ejemplo, el *British National Corpus* (BNC) o el COBUILD en inglés, y el Corpus *de referencia del español actual* (CREA) o el *Corpus del español del siglo XXI* (CORPES XXI) en español, a la enseñanza de lenguas (Wichmann *et al*. 1997; Burnard & McEnery 2000; Kettemann & Marko 2000; Aston 2001; Granger *et al*. 2002; Hunston 2002; Aijmer 2009), no se puede afirmar lo mismo en lo que concierne a las aplicaciones concretas de los corpus de aprendices o de estudiantes a la didáctica de segundas lenguas. Esto se debe a dos razones fundamentales: la compilación de corpus de aprendices ha sido, hasta la fecha, mucho más restringida y reciente, y además estos corpus no representan la lengua de referencia o lengua meta, sino uno o varios tipos de interlengua, lo cual también conlleva ciertas limitaciones.

En nuestro trabajo partimos del planteamiento general de Leech (1997: 5) quien, al referirse a las posibles aplicaciones de los corpus lingüísticos a la enseñanza de lenguas, establece una división clara entre los usos directos e indirectos, entendiendo por los primeros “the direct use of corpora as resources for teaching”, es decir, las aplicaciones de los corpus como recursos para la enseñanza y con un impacto concreto en la metodología de aula; en el caso de los

segundos, sin embargo, la contribución de los corpus a la didáctica de la lengua sería de carácter más secundario.

Adaptando este marco general a los corpus de aprendices en particular, que es lo que aquí nos ocupa, englobaríamos dentro de las aplicaciones directas las investigaciones realizadas sobre el aprendizaje del español/LE, la elaboración de actividades de aula y la confección de materiales didácticos. Como aplicaciones indirectas se encuadrarían aquellas que puedan tener una incidencia en la formación del profesorado, el diseño y desarrollo curricular, y en el ámbito de las pruebas lingüísticas y la evaluación. Nuestra exposición en las páginas que siguen estará organizada bajo estos epígrafes generales.

3.1 Investigación sobre el aprendizaje del español como lengua extranjera

Con cierta frecuencia, los profesores de lenguas nos mostramos más preocupados por el propio proceso de enseñanza que por el aprendizaje en sí, es decir, estamos más interesados en saber si una determinada técnica o actividad es realmente efectiva en el aula que por conocer detalles de cómo aprenden nuestros alumnos: qué dificultades se encuentran, cómo les gusta aprender, cuál es su motivación y su estilo de aprendizaje, qué estrategias ponen en marcha cuando no comprenden el significado de una palabra en un texto o no son capaces de descodificar un mensaje oral. Como resultado, no reparamos, a menudo, en que la información que podamos obtener sobre el proceso y experiencia de aprendizaje de nuestro alumnado podría ser de gran utilidad para iluminar, fortalecer e incluso mejorar determinados aspectos de nuestra docencia.

A tenor de esto, podemos distinguir tres grandes líneas de trabajo donde los corpus de aprendices, y el CAES en este caso, nos podrán proporcionar información de gran relevancia sobre el proceso de aprendizaje del español/LE de nuestro alumnado: (i) estudios comparativos entre el uso del español por parte de hablantes nativos y no nativos en la línea de lo que Granger (1998, 2002) definió como *Contrastive Interlanguage Analysis* (CIA), es decir, Análisis Contrastivo de Interlengua; (ii) investigaciones realizadas con el fin de confirmar o rechazar las hipótesis que apuntan a la existencia de una serie de áreas de la gramática española que presentan especial dificultad para el estudiante con el fin de aprovechar y revertir esta información en la propia enseñanza; (iii) estudios sobre la adquisición de morfemas gramaticales en el aprendizaje del español/LE en paralelo a los que se realizaron sobre el español como L1 y sobre otras lenguas como el inglés en la que este tipo de investigaciones adquirieron un gran auge.

3.1.1 El análisis contrastivo de muestras de interlengua

La comparación de producciones escritas y/u orales de hablantes nativos de español con las de estudiantes de español/LE nos podría reportar datos de interés con el fin de dilucidar si hay determinadas categorías gramaticales, estructuras sintácticas, expresiones o términos léxicos, colocaciones, intensificadores, mitigadores y marcadores discursivos que son más o menos utilizados por los miembros de un grupo que por los del otro. Resulta interesante saber en qué medida el uso del español por parte de nuestros alumnos difiere del propio de los hablantes habituales para así, de ser necesario, introducir ajustes en nuestra enseñanza. Bajo el marco del Análisis Contrastivo de Interlengua sería posible, por ejemplo, contrastar muestras del CAES correspondientes al nivel C1 con otras del CORPES XXI de una tipología similar, teniendo en cuenta que los textos escritos que conforman este último están clasificados por tipo y género textual. Es evidente que los resultados de investigaciones de esta naturaleza serían limitados puesto que estos dos corpus no son totalmente comparables; para que así fuera, sería necesario compilar dos corpus *ad hoc*, uno de hablantes nativos y otro de aprendices, que siguieran criterios similares en cuanto a su estructura, método de compilación, tamaño, tipología textual, perfil de los participantes, etc., tal como ya existe para el inglés con el *International Corpus of Learner English* (ICLE) y el *Louvain Corpus of Native English Essays* (LOCNESS), que consisten en colecciones de ensayos de universitarios sobre temas semejantes[1]. Sin duda, esta sería una tarea pendiente para el futuro. No obstante, estudios basados en estos dos corpus con los que ya contamos, CAES y CORPES XXI, podrían apuntar al menos tendencias valiosas en una u otra dirección al contrastar estas producciones lingüísticas. Sin embargo, más factible y metodológicamente más justificable sería realizar comparaciones entre las muestras de participantes del CAES de distintas lenguas maternas en torno a una pregunta o hipótesis de investigación, habida cuenta de que en el momento presente están representadas en este corpus producciones escritas de alumnos de al menos seis lenguas maternas diferentes, tal como se explicó más arriba. Esto nos permitiría investigar en qué medida los estudiantes con distintas L1 afrontan una misma cuestión de un modo similar o no. A modo de ejemplificación, veamos qué información podemos extraer del CAES cuando comparamos los ejemplos que aparecen con el verbo *gustar* en las muestras de estudiantes de francés y de inglés del mismo nivel, en este caso vamos a elegir el más bajo, es decir, A1. Seleccionamos para esta tarea muestras de estas dos lenguas por pertenecer a familias lingüísticas diferentes, lo que previsiblemente se concretará en resultados también diversos.

Para los estudiantes franceses de nivel A1, la herramienta de consulta recupera un total de 141 casos. Los datos de un primer análisis nos indican que para estos alumnos, a pesar de ser unos principiantes en el estudio del español, las construcciones con *gustar* no revisten, en

[1] Para más información sobre estos corpus se puede consultar los enlaces siguientes:
<https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html>
<https://www.learnercorpusassociation.org/resources/tools/locness-corpus/, ultimo acceso 12 de marzo de 2019.

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.

principio, una gran dificultad. Los mayores problemas se derivan de la falta de concordancia entre sujeto y verbo (*no me gusta los deportes*; *no me gusta los persones que no só simpática*; *no me gusta películas de horror*) pero globalmente utilizan este verbo con corrección, atreviéndose incluso con estructuras un tanto complejas, teniendo en cuenta el grado de dominio del español por parte de estos alumnos como, por ejemplo, *a mi me gustaría trabajar en vuestra cruadilla*; *le gustaría ser un profesor de ciencias*. Si vemos ahora qué ocurre con los estudiantes de habla inglesa del mismo nivel, observamos que el número de ejemplos es más reducido que en el caso anterior, un total de 107, detectándose un mayor número de formas no estándares derivadas, no solo de la falta de concordancia como en el caso anterior (*me gusta todos los deportes*), sino también del uso pronominal (*Mi abrina se llama Ruolin, solo tiene 15 meses, se me gusta mucho*), orden de palabras (*me gusta arroz mucho*) y conjugación verbal (*Me gusto juego el chess; me gusto salir de noche a un restaurante*). Sería interesante profundizar más en esta cuestión tratando de identificar en una segunda fase las razones que pudieran justificar estos resultados, bien debido a factores de transferencia lingüística o de otra naturaleza. Para ello sería preciso examinar de modo detallado las muestras de los distintos niveles de los dos grupos de hablantes fijando una serie de parámetros de carácter gramatical, léxico e incluso pragmático que nos proporcionaría una información más veraz y fiable. Todo esto nos ayudaría a comprender si determinados procesos de adquisición del español son universales o, de lo contrario, están condicionados por la lengua origen de los sujetos.

3.1.2 Problemas en el aprendizaje

Los datos proporcionados por este corpus nos servirán para confirmar o rechazar la hipótesis de que los alumnos de español/LE tienden a tener dificultades con el aprendizaje de determinadas áreas de la lengua a las que se suele prestar una atención especial en la mayor parte de las gramáticas pedagógicas, es decir, aquellos manuales gramaticales orientados de manera específica a la enseñanza de español/LE (Gómez, Pérez & Requeijo 1987; Benítez & Gelabert 1995; González Hermoso, Cuenot & Sánchez Alfaro 1999; Moreno 2001 Alonso Raya et al. 2005; Lieberman 2007; Areizaga Orube 2009; Palencia & Aragonés 2009; Romero Dueñas & González Hermoso, 2011; Garnacho & Martín, 2014; Villegas Galán & Blázquez Lozano 2014) o se identifican como tales en estudios globales de análisis de errores (Vázquez 1991; Santos Gargallo 1993, 2004). Entre estas áreas de especial dificultad destacan muchas que se pueden encuadrar dentro del nivel morfo-sintáctico de la lengua como son las diferencias entre los verbos *ser*, *estar* y *haber*, la conjugación y el uso del subjuntivo, los valores del indefinido frente al imperfecto, el uso de las preposiciones, en particular *para* frente a *por*, el género de los sustantivos, los artículos, la colocación de los pronombres, construcciones pasivas e impersonales con *se*, así como otras relacionadas más directamente con el léxico como los

falsos amigos, expresiones idiomáticas o determinadas colocaciones, e incluso otras en las que el nivel léxico y el morfosintáctico interactúan, como ocurre con las diferencias entre *recordar* y *acordarse*, construcciones con el verbo *gustar*, etc. Todo esto lo debemos concebir dentro de un gran objetivo final que no será otro que conseguir que el alumno adquiera una competencia comunicativa con lo que todo ello implica (Consejo de Europa 2002)[2]. Lógicamente se aboga por un estudio de la gramática como medio para la obtención de la comunicación en todas sus facetas y vertientes en la línea que apuntan varios autores en el ámbito concreto de la enseñanza del español/LE (Matte Bon 1992a, 1992b, 1998, 2004; Gómez del Estal Villarino 2004; Lieberman 2007: 19; Areizaga Orube 2009: 2). Este es un principio general de partida que debe estar siempre presente en nuestra enseñanza.

Sin profundizar ahora demasiado, sino a modo de esbozo y ejemplificación puesto que, tal como indicábamos más arriba, queremos animar al profesorado y especialistas en el uso del corpus, podemos ver en qué medida la confusión entre los verbos *ser* y *estar* resulta problemática para nuestros alumnos de español. Con este fin podemos llevar a cabo con la herramienta de consulta del CAES una búsqueda de la forma de primera persona del plural del presente del verbo *estar*, es decir, *estamos*. Una vez introducidos los datos en la aplicación del corpus, la herramienta de consulta recupera un total de 112 casos. Cuando analizamos estos resultados, observamos que en 20 de ellos, es decir, en alrededor del 18% del total, nos encontramos con un uso incorrecto y todos ellos corresponden, tal como por otra parte esperábamos, a muestras de estudiantes de niveles inferiores, A1, A2 e incluso B1, independientemente de su lengua origen, pues se presenta en alumnos cuya L1 es el árabe, el inglés o el chino, tal como muestran los ejemplos siguientes.

(1)

NIVEL	**L1**	**EJEMPLO**
A1	árabe	**Estamos** tres hermanos KHALID Y AHMED Y OFIANE[3].
A2	árabe	nosotros **estamos** amigos de 18 años

[2] Para el Consejo de Europa (2002: 13) la competencia comunicativa engloba tres componentes: el lingüístico, el sociolingüístico y el pragmático. Cada uno de esos componentes incluye, a su vez, una serie de conocimientos, destrezas y habilidades. Los conocimientos están relacionados simplemente con el saber qué, mientras que las destrezas y habilidades con el saber cómo. Las competencias lingüísticas comprenden, por su parte, los conocimientos y las destrezas léxicas, fonológicas y sintácticas, así como otras dimensiones de la lengua como sistema. Las competencias sociolingüísticas tienen que ver con las condiciones socioculturales del uso de la lengua mientras que las pragmáticas se refieren al uso funcional de los recursos lingüísticos sobre la base de contextos y escenarios de intercambios comunicativos. Este tipo de competencias engloban también todo lo relacionado con el dominio del discurso, la cohesión y la coherencia así como la identificación de los distintos tipos y formas de texto, la ironía y la parodia.

[3] Reproducimos los ejemplos seleccionados tal como se aparecen en el corpus sin introducir ningún cambio con respecto al original. En este caso los nombres propios aparecen todos ellos en mayúscula pues así los escribió el alumno. A este respecto es preciso tener en cuenta que en el CAES son los propios participantes los que introducen los datos de sus tareas escritas en la aplicación informática diseñada para tal propósito, evitándose de ese modo posibles errores en la transcripción derivados de falsas interpretaciones o de simples despistes del transcriptor.

A2	chino mandarín	Aquel entonces **estamos** felices cada día.
B1	inglés	**Estamos** amigas en Facebook y Tuenti también.

Así, la evidencia proporcionada por CAES confirma que efectivamente la diferencia entre *ser* y *estar* es problemática para los estudiantes de español y que, por lo tanto, el profesor debe prestarle una atención especial en el aula, realzando desde un punto de vista comunicativo cuándo se utiliza una forma u otra. Es evidente que el conocimiento de reglas de uso desde una perspectiva teórica no es suficiente puesto que es imprescindible saber llevar esas reglas a la práctica en la comunicación diaria. A tenor de esto y de los resultados anteriores, si profundizamos un poco más, observamos que cuando la forma investigada *estamos* va seguida de otra forma verbal, es decir, en construcciones durativas, los participantes del corpus no tienen en este caso mayores dificultades, como reflejan los siguientes ejemplos.

(2)

NIVEL	**L1**	**EJEMPLO**
A1	árabe	Yo soy guitarrista en una banda de música y **estamos** jugando le música jazz.
A1	portugués	**Estamos** haciendo cordeiro.
A2	portugués	**Estamos** verificando algunos sitios.
B1	francés	**Estamos** yendo a Bruselas por el trabajo.

Además de lo anterior, sería totalmente factible e incluso sencillo trasladar todos estos ejemplos al aula y discutirlos con nuestros alumnos en la línea de lo que presentaremos en la sección siguiente (3.2). En este sentido, el corpus no solo nos proporciona información de interés sino que también nos puede dar algunas claves para llevar esto a la clase.

Pasando ahora a otra área de conflicto potencial, el uso de las preposiciones, sabemos también por experiencia, al igual que en el caso anterior, que suele ser una fuente de problemas. A la luz de este punto de partida, podemos explorar en qué medida los datos extraídos del CAES sustentan esta hipótesis, intentando averiguar también como segundo objetivo qué preposiciones en particular plantean mayores dificultades y en qué casos concretos. Utilizando de nuevo la herramienta de consulta del CAES, observamos que la preposición *de* es la más frecuente en el corpus con un total de 18 240 ejemplos, seguida de las preposiciones *en* (13 661 ejemplos) y *a* (9545 unidades). Por el contrario, *tras*, *según* y *hacia* son las menos comunes con unas cifras mucho más bajas, 13 en el primer caso y 27 en el de las dos últimas. Además, un primer análisis de estos resultados nos indica que las preposiciones *para* y *por* resultan ser las que entrañan mayor dificultad; *desde* ocuparía un punto intermedio. Profundizando un poco más en estos datos, detectamos que de los cuatro significados de *para* que reseña de Bruyne (1999: 678-681), a saber, finalidad, movimiento, período de tiempo y relación de personas, cosas y situaciones del mismo tipo, es con el tercero de ellos, es decir, el de valor durativo, con el que

los estudiantes encuentran mayores dificultades al confundir con bastante frecuencia *para* con *por*. Los ejemplos siguientes tomados del corpus ilustran este hecho.

(3)

NIVEL	L1	EJEMPLO
A1	árabe	Pardon mi profesor, yo estoy en casa **para** tres días, porque yo soy muy enferma.
A2	portugués	¿Cuál es el precio total **para** 2 días?
A2	inglés	Yo vivi con mi papá y mi madrastra **para** dos meses.

Algo similar ocurre cuando *para* se utiliza para expresar dirección, movimiento.

(4)

NIVEL	L1	EJEMPLO
A1	portugués	Le gusta mucho su trabajo, pero también le gusta viajar **para** Alemania.
A2	portugués	Pero en Febrero , yo quiero viajar **para** Argentina.

Sin embargo, apreciamos también que con el significado de finalidad el número de confusiones es mucho menor. Esto es importante reseñarlo porque no solo resulta interesante centrarnos en las formas que no siguen la norma, sino también en los usos correctos, es decir, en aquello que hace bien el alumnado.

(5)

NIVEL	L1	EJEMPLO
A1	árabe	Yo leyó libos de gestión de empresas **para** amentar mi conocimiento.
A2	portugués	Voy a empezar mis estudios de derecho en la Universidad de Santiago de Compostela **para** ampliar mis conocimientos en la lengua española.

Si ahora llevamos a cabo un estudio similar con la preposición *por*, advertimos que los resultados del CAES nos confirman de nuevo nuestra pregunta de investigación puesto que los alumnos de todos los niveles tienen dificultades con su uso. Esta circunstancia puede derivarse de la multifuncionalidad de esta preposición; así, de Bruyne (1999: 681-690) reseña trece usos principales que van del valor agente en oraciones pasivas y duración a usos en oraciones concesivas (*por muy caro que sea*…) y exclamaciones (¡*por mis hijos que lo hago*!). Obsérvense los ejemplos siguientes:

(6)

NIVEL	L1	EJEMPLO
A1	árabe	El Mundo artístico es muy importante **por** mi.
A2	francés	Mi amigo Daniel tiene un problema con un pie y es un_poco dificile **por** el de caminar mucho.

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.

B1	chino mandarín	¿Podrias contar me más detalles sobre Tianjin y reservar un alojamiento **por** mi?
C1	árabe	Fue seleccionada **por** el festival_ de_ cannes, pero no tenia premio.

Las muestras recuperadas también nos indican que existe una tendencia a confundir *por* con *durante*.

(7)

NIVEL	**L1**	**EJEMPLO**
A1	Inglés	Era un profesor **por** más de treinta años.
B1	Inglés	He viajando con tu compañía **por** diez años[4].
B2	Árabe	Después de mi graduación viaje a España **por** dos semanas para asistir a un seminario de traducción Árabe.

Asimismo, se advierten confusiones cuando muchos de estos alumnos se refieren a medios de transporte, encontrándonos también con un número importante de colocaciones no gramaticales, tal como se puede apreciar en los ejemplos siguientes.

(8)

NIVEL	**L1**	**EJEMPLO**
A1	portugués	Soy aficionado **por** Real Madrid pero me gusta también el Barcelona.
A1	portugués	Soy responsable **por** todo el banco de dados de ventas de la compañía.
A2	inglés	Volveramos **por** autobús al centro de la cuidad.
A2	portugués	Mi familia y yo fuimos **por** carro y salimos muy temprano.
A2	chino mandarín	Está aficionado **por** películas y quieres ser un director en le futuro
B2	portugués	Tengo ganas **por** les conocer.

Al igual que en el caso anterior, es evidente que el profesor de español/LE tendrá que tener todo esto en cuenta en la planificación de sus clases y, sobre todo, a la hora de presentar y practicar estos contenidos gramaticales de la lengua.

3.1.3 Estudios sobre la adquisición de morfemas del español como lengua extranjera

En los años 70 y 80 del siglo pasado se realizaron, tomando como base el inglés como segunda lengua, una gran cantidad de investigaciones (Dulay & Burt 1973; Krashen & Terrel 1983; Krashen 1987) con el fin de averiguar si se podía hablar de un orden de adquisición similar o no al ya identificado para el inglés como L1, es decir, si en el proceso de aprendizaje de esta lengua se seguía una ruta semejante o no. Estos trabajos se basaban en el análisis de una serie de morfemas gramaticales tales como la *-s* de la tercera persona del presente simple, el uso

[4] En algunas variedades del español, este uso de *por* se considera totalmente gramatical.

de BE como cópula y en su valor progresivo, la formación del plural, el uso del genitivo con los nombres, los artículos, etc. La hipótesis que subyacía a estas investigaciones era que mediante el seguimiento longitudinal de un grupo de aprendices del inglés se podría llegar a la formulación de un orden de adquisición que supuestamente sería el mismo para todos los aprendices independientemente de su lengua materna. Esto reforzaría los presupuestos de la existencia de una Gramática Universal y pudiera tener ciertas implicaciones didácticas, ya que podría servir de base para el diseño curricular de programas de esta lengua (Pienemann 1989). Si bien estos estudios no estuvieron exentos de críticas (Hatch 1978; Ellis 2004), lo cierto es que efectivamente obtuvieron el éxito deseado puesto que fue posible llegar a establecer ese orden de adquisición para el inglés como L2 que, en líneas generales, resultó ser bastante semejante al orden de adquisición ya conocido e identificado previamente para el inglés como L1. Ya en fechas más recientes Housen (2002) realizó un estudio en el que investigaba la adquisición de morfemas verbales por parte de escolares norteamericanos de los cursos de tercero al undécimo, utilizando para ello muestras extraídas del *Corpus of Young Learner Interlanguage*, y, también en una línea semejante, Tono (2000) compara el orden de adquisición de morfemas gramaticales identificado a través de muestras japonesas de interlengua con el originalmente propuesto por Dulay & Burt (1973).

A la luz de estos resultados con referencia al inglés, Zobl & Liceras (1994) llevaron a cabo un trabajo en el que reconocen una secuencia de adquisición de morfemas en español/LE con cierto parecido a la del español nativo y en el que las marcas de presente indicativo, la negación con *no*, el presente durativo (*estoy hablando*), el futuro perifrástico (v*oy a hablar*) y el artículo indefinido (*un avión*) se adquieren antes, mientras que el uso de *ser* y *estar* como cópula, el pretérito perfecto y el imperfecto aparecen en estadios más avanzados del proceso (Baralo 1999). Estos resultados, en su conjunto, confirman los obtenidos por Van Naerssen (1980) en una investigación previa donde esta especialista también llegó a la conclusión de que los estudiantes americanos de español de su estudio mostraban tener pocos problemas a la hora de elegir el género del nombre cuando este aparecía modificado por un adjetivo. Sin embargo, la concordancia de género entrañaba para estos sujetos mayor complejidad que la concordancia de número. Estos dos resultados confirmaban, por otra parte, que no existía una gran diferencia entre el orden de adquisición del español como L1 y como L2. Caso contrario se evidencia con respecto al tiempo verbal, más concretamente en lo que se refiere a la diferencia entre el pretérito e imperfecto, donde sí se observan diferencias importantes entre el español como lengua nativa y lengua segunda. Además, Van Naerssen (1980: 153), a tenor de su análisis, sugiere la existencia de un estadio en el desarrollo del español como L1 y como L2 en el que los hablantes tienden a utilizar una forma verbal básica a la que posteriormente le añaden una serie de flexiones. Esta forma básica pudiera ser fácilmente la tercera persona del singular del

presente de indicativo o incluso el infinitivo, si bien es más probable que sea la primera que la segunda.

Dado que el CAES contiene muestras de alumnos cuyos niveles de competencia lingüística han sido muy controlados y que se pueden considerar fiables, sería posible también llevar a cabo trabajos similares a los anteriores, tomando como referencia los resultados previamente reseñados. Para ello sería aconsejable comenzar con el análisis de las muestras correspondientes a los alumnos de una lengua primera concreta para luego acometer estudios paralelos con las muestras de los alumnos de las distintas lenguas origen. Aun siendo conscientes de que el CAES no se puede considerar como un corpus longitudinal sino transversal ya que no se hace un seguimiento de la adquisición de los mismos aprendices a través del tiempo, el perfil académico de los participantes es bastante similar, lo que pudiera justificar un estudio de esta naturaleza, contribuyendo así a cubrir una de las lagunas existentes en la investigación de datos del español no nativo, tal como reconoce Liceras (1996: 239):

> Una de las grandes carencias con que nos enfrentamos a la hora de estudiar la gramática del español no nativo es la falta de estudios longitudinales que nos permitan analizar un corpus de datos suficiente.

3.2 Elaboración de actividades de aula con material del CAES

Si hasta el momento nos hemos venido refiriendo a aplicaciones del CAES centradas en el proceso de aprendizaje principalmente, en este apartado exponemos cómo se puede explotar este material de forma sencilla y clara en el aula. En esta línea partimos de que la aproximación pedagógica elegida estará centrada en torno al estudiante como principal protagonista del proceso enseñanza/aprendizaje (Nunan 1988a) y en el que el profesor adoptará el papel de guía, despertando la conciencia y sensibilidad gramatical y léxica del alumno en el uso del idioma en la línea de lo que Rutherford (1987) denomina “consciousness-raising” y “grammar awareness activities”. Podemos decir entonces que se trata de un aprendizaje por descubrimiento en el que la enseñanza de la lengua se hace de forma inductiva y los alumnos van progresando en su aprendizaje bajo la supervisión del profesor como si fueran pequeños investigadores de la lengua. Este tipo de acercamiento didáctico, que toma como recurso fundamental los datos proporcionados por un corpus, es lo que se ha venido en llamar en inglés *corpus/data-driven learning*, es decir, aprendizaje derivado de los datos de un corpus. Los trabajos en esta línea son numerosos (Tribble & Jones 1990; Johns 1991; Granger & Tribble 1998; Tribble 2015) y todos ellos inciden en los grandes beneficios que puede reportar la explotación de los corpus, ya sean generales o de aprendices, para la enseñanza de lenguas.

La primera actividad que proponemos, dirigida especialmente a estudiantes de los niveles iniciales aunque adaptable a otros más avanzados, se centra en el significado, uso y ortografía de la conjunción causal *porque*, tomando como base muestras del CAES producidas por alumnos del nivel A2. Nuestro propósito último es que el alumno sea capaz de expresar causalidad tanto oralmente como por escrito y lo sepa hacer con corrección y de acuerdo con el contexto. Hemos elegido este punto gramatical porque es un tema recurrente en los manuales de enseñanza del español, tiene que ver con varios ámbitos de la lengua (sintaxis, semántica, ortografía, discurso, etc.), circunstancia que lo hace especialmente interesante, y además hemos podido apreciar a través del análisis del material del corpus que plantea dificultades en su aprendizaje. En nuestra propuesta partimos de la presentación de datos para que el discente observe el funcionamiento de la lengua y vaya llegando a sus propias conclusiones. A medida que avanza la actividad, se incrementa también su grado de dificultad hacia un uso de la lengua más autónomo y creativo.

Para comenzar, presentamos a los alumnos los ejemplos siguientes para que los lean detenidamente.

1. Estoy muy bien porque ayer volví a mi casa de Hanolulu, Hawaii.
2. Es un hombre sincero muy generoso porque ayuda a la gente.
3. No me acuerdo del nombre y tampoco de la historia porque dormí todo el tiempo.
4. Le admiro porque él es honesto, divertido, responsable y un amigo muy bueno.
5. La admiro porque ella es simpática.

A continuación, les formulamos las preguntas siguientes:

1. ¿Qué palabra se repite en todas las oraciones anteriores? Subráyala, por favor.
2. Empareja los elementos de las 2 columnas de modo que tengan sentido.

Ejemplo: 1 e. Admiro a Mike **porque** es el hombre más simpático del mundo.

1. Admiro a Mike	a. porque es el verano aquí.
2. Hace mucho calor	b. porque es muy bonita y grande
3. Me gusta pasar tiempo con él	c. porque es un regalo de mi esposo.
4. Me ha encantado Barcelona	d. porque hizo mucho calor
5. La maleta es muy importante para mí	e. porque es el hombre más simpático del mundo.
Fuimos a la playa todos los días	f. porque es muy divertido.

3. ¿Qué significado expresa? Elige la respuesta correcta:

a. tiempo

b. causa

c. condición

d. lugar

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.

e. fin

4. Ahora presta atención a su ortografía, es decir, a cómo se escribe ¿En qué casos se puede escribir como dos palabras separadas, es decir, *por qué*, como una palabra con tilde *porqué* y como una palabra sin tilde *porque*? Observa los ejemplos siguientes:

   No entiendo *por qué* te enfadas.

   ¿*Por qué* no hiciste los deberes? *Porque* no tuve tiempo.

   No saben el *porqué* de su comportamiento.

   ¿Cuáles son tus conclusiones?

   ……

   ……

5. Lee las oraciones siguientes e identifica los errores que encuentres. A continuación, escribe la oración correcta.

   Ejemplo: Voy/llegar más tarde en la casa por que he quedado con amigos en un bar de tapas.

   Voy a llegar más tarde a casa porque he quedado con amigos en un bar de tapas.

   1. Voy llegar más tarde en la casa por que he quedado con amigos en un bar de tapas.
   2. La visitaba y la preguntaba porque ha hecho esto.
   3. No es bueno por la salud y tampoco por el bolsillo porque las cigaretas no son baratas.
   4. Espero que tiene ya habitaciones libres por que es la alta período.
   5. Le pregunte porque, pues havia lo suficiente por 40 personas.

6. Completa las oraciones siguientes utilizando una de las formas estudiadas:

   1. No puedo opinar …….
   2. ¿Por qué estás triste hoy? …….
   3. Me cae bien Pedro ……..
   4. Prefiero callarme ……
   5. ¿Por qué llegaste tarde? ……..

La segunda tarea que proponemos dentro de esta sección está centrada en la enseñanza de la ortografía. Las muestras del CAES reflejan la dificultad que entrañan palabras con el grupo inicial *dif*, tales como "diferencia", "diferente", "difícil", "dificultad", ya que los alumnos tienden a escribirlas *diff*, posiblemente por transferencia del inglés. De hecho, nos encontramos

con 38 casos de estas características. Ocurre algo similar con palabras con el grupo inicial *col* (*colaboración*, *colegio*, *colega*) que tienden a escribirlas con doble l, *collega*, *collaboración*, *collegio*, etc. Identificamos 15 ejemplos. Lo mismo pasa con palabras que en español contienen el grupo *cc* (*acción*, *construcción*, *dirección*, *ficción*, etc.) y que las suelen escribir con una *c* nada más.

Al igual que en la actividad anterior, les presentamos ejemplos donde se utilizan las palabras seleccionadas para que extraigan sus propias conclusiones y donde llamamos la atención sobre la ortografía de estos términos.

No noto mucho la di**f**erencia	Hablaba con di**f**icultad.
Me encanta cocinar di**f**erentes comidas.	La situación es di**f**ícil

Muchas gracias por tu co**l**aboración.
El co**l**egio de mis hijos es muy grande.
Le gusta mucho comer con sus co**l**egas de trabajo.

Esta es la dire**cc**ión de la casa.
Es un edificio de constru**cc**ión moderna.
La película es de ciencia fi**cc**ión.

A continuación, les proponemos la siguiente tarea:

diferente/diferir	*coleccionar/ coleccionista*	*protección*
dificultad	*acción*	*sección*
colegio	*construcción*	*introducción*
colega (el, la)	*ficción*	*colección*
colaborar/ colaborador	*elección*	*reacción*

Completa las oraciones siguientes con una de las palabras dadas en los recuadros superiores. Si tienes alguna duda sobre su significado, consulta el diccionario.

1. Mi amiga trabaja como ……. en el periódico local.
2. La …….. del presidente se pospone hasta mañana.
3. Tiene una …….. de monedas muy interesante.
4. La …….. de la nueva autopista se realizará en 2020.
5. La ..... de datos es importante hoy en día.

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.

6. Siento ……. de tu opinión.

7. Debemos ponernos en …….

3.3 Elaboración de materiales didácticos

Los corpus de aprendices nos pueden proporcionar datos de interés que se pueden incorporar con cierta facilidad en gramáticas pedagógicas, diccionarios, glosarios y libros de texto concebidos específicamente para el aprendizaje del español/LE (Véase apartado 3.1.2). Así, por ejemplo, en los diccionarios y glosarios, se podría llamar la atención a sus usuarios del término en español que están buscando con respecto a otra palabra paralela de su L1 que presenta una ortografía semejante pero que, sin embargo, posee un significado total o parcialmente diferente, lo que se conoce como "falsos amigos" (Chacón Beltrán 2006; Chamizo Domínguez 2008; Roca Varela 2015[5]. De los datos del CAES se deriva que los estudiantes de español cuya lengua materna es el inglés tienden a identificar como semejantes, entre otros, los pares de palabras *suburb/suburbio*, *idiom/idioma*, *firm/compañía* mientras que los de habla francesa tienen el mismo problema con *champagne/campiña*, *sentiment/impresión*, y los alumnos cuya L1 es el portugués con los pares *aula/clase*, *brincar/bromear*, *polvo/pulpo*, *romance/novela*; incluso hay términos que llevan a confusión a alumnos de distintas lenguas maternas, por ejemplo, *aplicar* con el sentido de solicitar o *largo* expresando gran tamaño. Además de esto, también se puede advertir al alumno sobre ciertas características peculiares del término en cuestión referidas a su naturaleza gramatical o a su uso. Así, a modo de ilustración, nuestras búsquedas en el CAES nos revelan que un número considerable de alumnos, sobre todo aquellos que tienen el portugués como L1, utilizan la combinación *ir* + infinitivo en lugar de la perífrasis *ir* + a + infinitivo, por ejemplo, *Julia fue comprar algo* en lugar de *Julia fue* ***a*** *comprar algo*; *fue conocer la Normandía* en lugar de *fue* ***a*** *conocer Normandía*. Detectamos también que existe una tendencia a designar los nombres de los países o continentes con el artículo determinado, y así se refieren a *la Libia*, *la Europa*, *la Francia*, *la Lituania*, *el Egipto* como en los ejemplos siguientes: *Después fue a la Francia*; *Va a volver a el Egipto*; *voy a ir a el España*; *mi país de nacionamento es la Colombia*; *prefiero el Marocco*. Lo mismo ocurre con la distinción entre *saber* y *conocer* (*Conoce español pero le gustaría seguir un curso de cocina*), las diferencias entre *qué* y *cuál* (*¿qué son las condiciones de admisión*?), cuestiones relacionadas con el número del substantivo (*quiero saber más informaciones*), la colocación de los pronombres en la cláusula (*soy muy impaciente de conocer a vosotros*; *tengo ganas de conocer a vosotros*; *espero os conocer*). En realidad, sería posible añadir muchos más temas a

[5] Véase un trabajo anterior (AUTORES, 2016) donde analizábamos este tema con cierto detalle.

este listado puesto que a medida que se va buceando en el corpus, más información de relevancia pedagógica se encuentra.

3.4 Formación del profesorado

Las aplicaciones de un corpus como el CAES a la formación del profesorado pueden ser de carácter general o más particular. Así, por ejemplo, se podría interpretar que algunas de las carencias detectadas en las muestras de los estudiantes en su uso del español pudieran tener su origen en debilidades o lagunas en la formación lingüística y pedagógica del profesorado. En ocasiones, cuando se analizan los programas de cursos y seminarios de formación docente de español/LE así como de otras lenguas, observamos que se suele poner el acento, entre otros, en teorías y aproximaciones didácticas, temas de innovación educativa y aplicaciones de las nuevas tecnologías sin abordar problemas concretos con los que el profesorado se tendrá que enfrentar en su día a día en el aula. En esta línea, sería posible entonces presentar a los docentes algunos de los ejemplos anteriores donde se advertían dificultades (*ser* frente a *estar*, preposiciones, uso del subjuntivo, colocación de los pronombres, falsos amigos, etc.) y tratar con ellos su tratamiento didáctico, incluyendo aquí actividades y ejercicios que se pudieran realizar con los alumnos en función de su nivel y de su perfil personal y académico. Asimismo, material de este tipo podría utilizarse muy fácilmente a la hora de abordar la corrección de errores y el tipo de comentarios que se deben emitir sobre la producción oral y escrita de los estudiantes, y sobre la calidad de sus trabajos. Sería incluso posible seleccionar muestras de los distintos niveles y solicitar al profesorado participante en un curso o seminario que clasificase esas muestras en diferentes niveles de dominio de la lengua (A1, A2, B1, etc.), tomando como base una serie de criterios definidos previamente. A todo lo anterior, habría que añadir que no estaría de más que se incluyese dentro de la formación del profesorado de lenguas extranjeras un pequeño módulo cuyo propósito sería la familiarización del profesorado con los corpus y su explotación didáctica, ya no solo como recurso y herramienta didáctica para sus clases sino también para su propio desarrollo profesional. Es evidente que a través de los corpus se pueden contrastar y aprender muchas cuestiones sobre el funcionamiento y uso de la lengua: frecuencias, colocaciones, registro, variación, actitudes de los hablantes, variables lingüísticas y extralingüísticas, etc.

3.5 Diseño y desarrollo curricular

Por regla general, los programas de los cursos de español/LE, al igual que ocurre con los programas de otras lenguas extranjeras, son diseñados por las autoridades educativas correspondientes, tomando como base los habituales principios y prácticas curriculares

establecidos por especialistas en este campo (Nunan 1988b; Johnson 1989; Richards 2001; Medgyes & Nikolov 2010), así como documentos de referencia como el *Marco Común Europeo*, diccionarios y bibliografía especializada. Este es el caso del *Plan Curricular del Instituto Cervantes,* que fue elaborado por la Dirección Académica del Instituto Cervantes con la colaboración de un grupo de profesores y expertos, y los programas de las Escuelas Oficiales de Idiomas confeccionados por la administración educativa, es decir, el Ministerio de Educación y las Consejerías de Educación de las Comunidades Autónomas. Una vez que estos programas son publicados y puestos al servicio de la comunidad educativa, el profesorado elabora sus programaciones docentes tratando de adaptar los diseños curriculares oficiales a las especifidades y necesidades de su centro y de su alumnado. Los libros de texto también juegan un papel importante a este respecto porque, con frecuencia, condicionan las programaciones del profesorado pudiéndose convertir de facto en los verdaderos programas.

En función de lo anterior, el CAES podría utilizarse como material de referencia curricular pues contiene información real, ordenada y organizada por niveles que refleja el dominio del español, al menos en cuanto a su producción e interacción escrita, por parte de los alumnos que realizaron las tareas propuestas. Dado que la mayor parte de las muestras del corpus corresponden a alumnado de distintos centros del Instituto Cervantes de todo el mundo, sería factible investigar, por ejemplo, en qué medida existe una correspondencia entre la producción lingüística de estos alumnos con los niveles del *Marco Europeo* y con gran parte de los inventarios que se establecen en el *Plan Curricular del Instituto Cervantes* a propósito de la gramática, léxico, ortografía, técnicas y estrategias pragmáticas, etc. Es evidente que el proceso sería laborioso y que no sería posible cubrir todos y cada uno de los campos, pero sí serviría para proporcionar información necesaria para hacer un seguimiento de las actuales propuestas curriculares y tener así la base para la formulación de una serie de pautas de mejora. Este proceso parece pertinente pues se puede correr el peligro de que los diseños curriculares y, con ellos, los programas para cada uno de los niveles se queden en simples abstracciones incluyendo solamente aquello que se considera deseable pero que no resulta alcanzable en el día a día.

3.6 Otras posibles aplicaciones: evaluación, elaboración de pruebas

En los últimos años son más las voces (Mukherjee 2009; Barker 2010; Callies et al, 2014; Callies & Götz 2015) que llaman la atención sobre la utilidad de los corpus de aprendices para el proceso de evaluación de la lengua extranjera. Así, por ejemplo, Callies & Götz (2015:3) se refieren a esta cuestión en los términos siguientes:

> Generally speaking, learner corpora have the potential to increase transparency, consistency and comparability in the assessment of L2 proficiency, and in particular to inform, validate and advance the way L2 proficiency is assessed in the CEFR.

En este sentido el CAES, al recoger muestras representativas y contrastadas de cada nivel de competencia lingüística, nos proporciona información veraz sobre lo que son capaces de hacer los alumnos de cada nivel al mismo tiempo que nos revela sus limitaciones y dificultades. De este modo podría servir para aumentar la transparencia y solidez en la evaluación del español/LE. Sería factible también analizar las certificaciones existentes del español/LE y examinar en qué medida se adecúan a los diferentes grados de dominio del alumnado, es decir, investigar hasta qué punto existe una correspondencia entre las actividades propuestas en estas pruebas y la competencia lingüística esperable de los alumnos de cada nivel. Además de esto, y entrando ya en el ámbito de la Lingüística computacional, material del CAES podría utilizarse también para el diseño e incluso la verificación de la efectividad de determinadas herramientas o instrumentos de detección de errores que de forma automática los clasifican de acuerdo con una serie de categorías fijadas de antemano (sintaxis, léxico, ortografía). Una experiencia de este tipo se ha realizado ya con dos corpus de aprendices de gallego con unos buenos resultados (Gamallo *et al.* 2015). En una línea similar, el corpus en parte o en su totalidad pudiera también constituir un buen recurso para la creación de una base de datos útil para la confección de pruebas y exámenes.

BIBLIOGRAFÍA

Aijmer, Karin (ed.) (2009): *Corpora and Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Alonso Raya, Rosario, Alejandro Castañeda Castro, Pablo Martínez Gila, Lourdes Miquel López, Jenaro Ortega Olivares & José Plácido Ruiz Campillo (2005): *Gramática básica del estudiante de español.* Barcelona: Difusión.

Areizaga Orube, Elisabet (2009): *Gramática para Profesores de Español como Lengua Extranjera*. Madrid: Ediciones Díaz de Santos.

Aston, Guy (ed.) (2001): *Learning with Corpora*. Houston: Athelstan.

Baralo, Marta (1999): *La adquisición del español como lengua extranjera*. Madrid: Arco Libros.

Barker, Fiona (2010): "How can corpora be used in language testing?", en A. O'Keeffe & M. McCarthy (eds.): *The Routledge Handbook of Corpus Linguistics*. New York: Routledge, pp. 633-645.

Benítez, Pedro & María José Gelabert (1995): *Breve gramática Español lengua extranjera*. Barcelona: Difusión.

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.

Burnard, Lou & Tony McEnery (eds.) (2000): *Rethinking Language Pedagogy from a Corpus Perspective*. New York: Peter Lang

Callies, Marcus, María Belén Diez-Bedmar & Ekaterina Zaytseva (2014): "Using learner corpora for testing and assessing L2 proficiency", en P. Leclercq, H. Hilton & A. Edmonds (eds.): *Measuring L2 Proficiency: Perspectives from SLA* (Second Language Acquisition series). Clevedon: Multilingual Matters, pp. 71-90.

Callies, Marcus & Sandra Götz (eds.) (2015): *Learner Corpora in language Testing and Assessment*. Amsterdam/Philadelphia: John Benjamins.

Chamizo Domínguez, Pedro (2008): *Semantics and Pragmatics of False Friends*. New York: Routledge.

Chacón Beltrán, Rubén (2006). "Towards a typological classification of false friends (Spanish-English)", *Revista Española de Lingüística Aplicada* 19, pp. 29-39.

Consejo de Europa (2002). *Marco Común Europeo de Referencia para las lenguas: Aprendizaje, Enseñanza, Evaluación*. Madrid: Ministerio de Educación, Cultura y Deporte.

De Bruyne, Jacques (1999): "Las preposiciones", en Bosque, I. & V. Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa, pp. 657-704.

Dulay, Heidi C. & Marina K. Burt (1973): "Should we teach children syntax?", *Language Learning*, 23, pp. 245-258.

Ellis, Rod (2004): *Understanding Second language Acquisition*. Oxford: Oxford University Press.

Gamallo, Pablo, Marcos García, Iria del Río & Isaac González (2015): "Natural language processing for automatic error detection", en M. Callies & S. Götz (eds.): *Learner Corpora in Language Testing and Assessment*. Amsterdam/Philadelphia, John Benjamins, pp. 35-57.

Garnacho López, Pilar & María Dolores Martín Acosta (2014): *Diccionario de dudas del estudiante de español como lengua extranjera*. Madrid: SGEL.

Gómez, María Dolores, María Jesús Pérez & María H. Requeijo (1987): *Gramática española para extranjeros*. Santiago de Compostela. Servicio de Publicaciones de la Universidad.

Gómez del Estal Villarino, Mario (2004): "Los contenidos lingüísticos o gramaticales. La reflexión sobre la lengua en el aula de E/LE: Criterios pedagógicos, lingüísticos y psicolingüísticos", en J. Sánchez Lobato & I. Santos Gargallo (dirs.): *Vademécum para la formación de profesores. Enseñar español como lengua segunda (L2)/Lengua extranjera (LE)*. Madrid: SGEL, pp. 767-787.

González Hermoso, Alfredo, J. R. Cuenot & María Sánchez Alfaro (1999): *Gramática del español lengua extranjera*. Madrid: Edelsa.

Granger, Sylviane (ed.) (1998): *Learner English on Computer*. Longman: New York.

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.

Granger, Sylviane (2002): “A bird’s eye view of learner corpus research”, en S. Granger, J. Hung & E. Petch-Tyson (eds.): *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, New York: Longman, pp. 3-33.

Granger, Sylviane, Joseph Hung & Stephanie Petch-Tyson (eds.) (2002): *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Philadelphia. John Benjamins

Granger, Sylviane & Chris Tribble (1998): “Exploiting learner corpus data in the classroom: form-focused instruction and data-driven learning”, en S. Granger (ed.): *Learner English on Computer*. New York, Longman, pp. 199-209.

Hatch, Evelyn M. (ed.) (1978): *Second language Acquisition. A Book of Readings*. Rowley, Mass: Newbury House Publishers.

Housen, Alex. (2002): “A corpus-based study of the L2-acquisition of the English verb system”, en S. Granger, J. Hung & S. Petch-Tyson (eds.): *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, pp. 77-116.

Hunston, Susan (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Instituto Cervantes. *Plan Curricular*.
<https://cvc.cervantes.es/ENSENANZA/biblioteca_ele/plan_curricular/default.htm, (última consulta, 12/03/2019)>

Johns, Tim (1991): “Should you be persuaded: Two examples of data-driven learning”, en T. Johns & P. King (eds.): *Classroom Concordancing. English Language Research Journal*, 4, pp. 1-16.

Johnson, Robert K. (1989): *The Second Language Curriculum*. Cambridge: Cambridge University Press.

Kettermann, Bernhard & Georg Marko (eds.) (2000): *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora*. New York: Rodopi.

Krashen, Stephen (1987): *Principles and Practice in Second Language Acquisition*. Englewood Cliffs: Prentice Hall.

Krashen, Stephen & Tracy D. Terrel (1983): *The Natural Approach: Language Acquisition in the Classroom*. Oxford: Pergamon.

Leech, Geoffrey (1997) “Teaching and language corpora: a convergence”, en A. Wichmann, S. Fligelstone, A. McEnery & G. Knowles (eds.): *Teaching and Language Corpora*. London: Longman, pp. 1-23.

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.

Liceras, Juana M. (1996): *La adquisición de las lenguas segundas y la gramática universal*. Madrid: Síntesis.

Lieberman, Dorotea I. (2007): *Temas de gramática del español como lengua extranjera*. *Una aproximación pedagógica*. Buenos Aires: Editorial Universitaria de Buenos Aires.

Matte Bon, Francisco (1992a): *Gramática comunicativa del español. De la lengua la idea, I*. Barcelona: Difusión.

Matte Bon, Francisco (1992b): *Gramática comunicativa del español. De la lengua la idea, II*. Barcelona: Difusión.

Matte Bon, Francisco. (1998): "Gramática, pragmática y enseñanza comunicativa del español", *Carabela*, 43, pp. 53-79.

Matte Bon, Francisco (2004). "Los contenidos funcionales y comunicativos", en J. Sánchez Lobato & I. Santos Gargallo (dirs.): *Vademécum para la formación de profesores. Enseñar español como lengua segunda (L2)/Lengua extranjera (LE)*. Madrid: SGEL, pp. 811-834.

Medgyes, Péter & Marianne Nikolov (2010): "Curriculum development in foreign language education: The interface between political and professional education", en R. K. Kaplan (ed.): *The Oxford Handbook of Applied Linguistics*. Oxford: Oxford University Press, pp. 264-274.

Moreno, Concha (2001): *Temas de gramática*. Madrid: SGEL.

Mukherjee, Joybrato (2009): "The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications", en K. Aijmer (ed.): *Corpora and Language Teaching*. Amsterdam: John Benjamins, pp. 203-230.

Nunan, David (1988a): *The Learner-centred Curriculum*. Cambridge: Cambridge University Press.

Nunan, David (1988b): *Syllabus Design*. Oxford: Oxford University Press.

Palencia, Ramón & Luis Aragonés (2009): *Gramática y uso del español para extranjeros. Teoría y Práctica* (distintos niveles). Madrid: S.M.

Pienemann, Michael (1989): "Is language teachable? Psycholinguistic experiments and hypothesis", *Applied Linguistics*, 10, pp. 52-79.

Richards, Jack C. (2001): *Curriculum Development in Language Teaching*. Cambridge: Cambridge University Press.

Roca Varela, María Luisa (2015): *False Friends in Learner Corpora***.** *A Corpus-based Study of English False Friends in the Written and Spoken Production of Spanish Learners*. Berna: Peter Lang.

AUTORES (2016): "Learner Spanish on Computer. The CAES 'Corpus de Aprendices de Español' Project", en M. Alonso-Ramos (ed.): *Spanish Learner Corpus Research. Current Trends and Future Perspectives*. Amsterdam: John Benjamins, pp. 55-86.

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.

Romero Dueñas, Carlos & Alfredo González Hermoso (2011): *Gramática del español lengua extranjera*. Madrid: Edelsa.

Rutherford, William (1987): *Second Language Grammar: Learning and Teaching*. London: Routledge.

Santos Gargallo, Isabel (1993): *Análisis contrastivo, análisis de errores e interlengua en el marco de la lingüística contrastiva*. Madrid: Síntesis.

Santos Gargallo, Isabel (2004): "El análisis de errores en la interlengua del hablante no nativo", en J. Sánchez Lobato & I. Santos Gargallo (dirs.): *Vademécum para la formación de profesores. Enseñar español como lengua segunda (L2)/Lengua extranjera (LE)*. Madrid: SGEL, pp. 391-410.

Tono, Yukio (2000): "A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes", en L. Burnard & T. McEnery (eds.): *Rethinking Language Pedagogy from a Corpus Perspective*. New York: Peter Lang, pp. 123-132.

Tribble, Chris (2015): "Teaching and language corpora: Perspectives from a pedagogical journey", en A. Leńko-Szymańska & A. Boulton (eds.): *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins, pp. 37-62.

Tribble, Chris. & Glyn Jones (1990): *Concordances in the Classroom*. Harlow: Longman.

Van Naerssen, M. M. (1980): "How similar are Spanish as a first language and Spanish as a foreign language?", en R. Scarcella & S. Krashen (eds.): *Research in Second Language Acquisition. Selected Papers of the Angeles Second Language*. Rowley, Mass: Newbury House Publishers, pp. 146-154.

Vázquez, Graciela E. (1991): *Análisis de errores y aprendizaje de español/lengua extranjera*. Frankfurt am Main: Peter Lang.

Villegas Galán, María de los Angeles & María Jesús Blázquez Lozano (2014): *Universo gramatical. Gramática de referencia para estudiantes de español*. Madrid: Edinumen.

Wichman, Anne., Steve Fligelstone, Tony McEnery & Gerry Knowles (1997): *Teaching and Language Corpora*. Londres y New York: Longman.

Zobl, Helmut & Juana Liceras (1994): "Functional categories and acquisition orders." *Language Learning*, 44, 1, pp. 159-180.

Borrador final. Pendiente de publicación en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, 2019, 273-303.