CORPUS PARA EL ANÁLISIS DE ERRORES DE APRENDICES DE E/LE (*CORANE)*

Ana Mª Cestero Mancera, Inmaculada Penadés Martínez, Ana Blanco Canales,
Laura Camargo Fernández y José Simón Granda
Universidad de Alcalá

1. INTRODUCCIÓN

La Lingüística Aplicada a la enseñanza y el aprendizaje de lenguas extranjeras es un ámbito de estudio ya consolidado, que ha tenido un rápido desarrollo y que tiene un auge creciente en la actualidad, sobre todo en este Estado. Son ya muchos los proyectos de investigación, las tesis doctorales y las monografías realizados con el fin de profundizar en el conocimiento de los procedimientos para la enseñanza y la adquisición de lenguas extranjeras.

Muchos de los estudios en este campo, especialmente los relativos a análisis de errores e interlengua, han de hacerse a partir de un corpus de materiales producidos por estudiantes extranjeros, de forma que se trabaje sobre datos reales recogidos de forma sistemática y organizada. La elaboración de un amplio corpus de estas características es, sin embargo, una tarea costosa en tiempo y esfuerzo, por lo que, hasta ahora, los investigadores han elaborado *corpora* reducidos que les permitieran realizar trabajos puntuales atendiendo a los objetivos de su propia investigación. Así, no existe en la actualidad para el español, al menos hasta donde llega nuestra información, ningún corpus extenso, de grandes dimensiones[1], que sea de fácil acceso y que permita llevar a cabo un amplio número de investigaciones diferentes, enmarcadas en el ámbito de la lingüística aplicada a la enseñanza y adquisición de E/LE, tales como: análisis de errores lingüísticos; problemas de adquisición relacionados con la lengua materna, con el nivel de conocimientos, con el sexo o con la edad de los estudiantes; tipos, frecuencias, condicionamientos e implicaciones de las transferencias lingüísticas; elaboración de materiales didácticos para la enseñanza de la gramática o del léxico, etc.[2]

La enseñanza del español como lengua extranjera es un campo de investigación y una línea de especialización fundamentales en la Universidad de Alcalá, como demuestran la amplia programación desarrollada en los Cursos de Lengua y Cultura Españolas para Extranjeros, organizados e impartidos en esta Universidad, y la excelente aceptación que

[1]. Sí nos consta, no obstante, su existencia para el inglés. Así, es obligado mencionar aquí el *International Corpus of Learner English* (ICLE) dirigido por Sylviane Granger.

[2]. Tenemos noticia, por una comunicación publicada en las *Actas del X Congreso Internacional de ASELE*, de que Fermín Sierra Martínez (2000) ha recogido un corpus de errores producidos por alumnos de Filología Hispánica de la Universidad de Amsterdam.

tiene el *Máster en Enseñanza de Español como Lengua Extranjera* desde que se impartió por primera vez en el curso 1994-95. Sin embargo, para el desarrollo de una y otra tarea es fundamental el conocimiento de los errores más frecuentes y más característicos de los alumnos que aprenden español, pues sólo así el profesor podrá conocer de antemano las principales áreas de dificultad a las que se enfrentarán sus alumnos y se podrán preparar materiales didácticos adecuados a cada grupo de aprendices, en función de su lengua materna. Ahora bien, el análisis de errores cometidos por los estudiantes extranjeros que aprenden español es un campo de investigación que todavía no ha alcanzado en España el desarrollo y la amplitud que serían deseables, entre otras razones tal vez por la no existencia de *corpora* lingüísticos que permitan analizar de manera sistemática los errores de estos estudiantes. Así, por ejemplo, esta carencia ha dificultado enormemente, en los últimos años, tanto nuestros propios estudios sobre análisis de errores y nuestras publicaciones de materiales didácticos[3], como las investigaciones que nuestros estudiantes han realizado para completar su máster; por ello, después de varias tentativas[4], decidimos, a finales de 1999, acometer la empresa de llenar el vacío existente con la creación de un corpus de materiales escritos recogido de forma sistemática durante el año 2000, en los Cursos de Lengua y Cultura Españolas para Extranjeros de la Universidad de Alcalá. Dicho proyecto, financiado por el Consejo Social de la mencionada Universidad[5], es el que aquí presentamos.

2. El corpus de materiales escritos de la Universidad de Alcalá

Los objetivos generales del proyecto que reseñamos son dos: la creación y preparación de un corpus para el estudio de la enseñanza-aprendizaje de E/LE a partir de textos escritos por los alumnos matriculados en los distintos Cursos de Lengua y Cultura Españolas para Extranjeros de la Universidad de Alcalá durante el año 2000, en una primera fase, y la informatización, el almacenamiento y el etiquetado del corpus recogido, diferenciando en él los textos de acuerdo con la lengua materna del estudiante, su nivel de conocimiento de español y otras variables sociales que, en principio, puedan resultar

[3]. Véase Álvarez Martínez, Blanco Canales, Gómez Sacristán y Pérez de la Cruz (2001); Blanco Canales, Fernández López y Torrens Álvarez (2001); Cabrerizo Ruiz, Gómez Sacristán y Ruiz (2001); Fuente, Giraldo Silverio, Martín Martín , Sanz Sánchez y Torrens Álvarez (2001), y Morimoto y Penadés (2001).

[4]. Nos referimos, concretamente, a los *corpora*, todos de reducidas dimensiones, que se han recogido para realizar distintas memorias de investigación para el *Máster en Enseñanza de Español como Lengua Extranjera.* Estas monografías tienen como tema central el análisis de errores de estudiantes de español cuya lengua materna es el portugués de Brasil (Duarte 1999; Tomazini 1997; Mansilla Clemente 1999, y Damas Gil 2000), el chino hablado en Taiwan (Lin 1998), el chino hablado en la República Popular China (Li 1999), el inglés hablado en E.E.U.U. (Aznar Juan 1998) y el italiano (Gutiérrez Quintana 2000). Una parte de los resultados obtenidos en estas investigaciones está expuesta en Penadés Martínez (1999).

[5]. Se trata del proyecto H011/2000, subvencionado por la Universidad de Alcalá, titulado: "Corpus de materiales escritos para el estudio de la enseñanza y el aprendizaje del español como lengua extranjera", en el que participan los siguientes investigadores: Ana Mª Cestero Mancera, investigadora principal, Fátima Álvarez López, Ana Blanco Canales, Laura Camargo Fernández, Inmaculada Penadés Martínez, Ana Mª Ruiz Martínez, José Simón Granda y Mª Jesús Torrens Álvarez.

factores condicionantes del aprendizaje de la lengua española, como la edad y el sexo de los estudiantes. Dichos objetivos, ya cumplidos, nos permitirán realizar un trabajo de mayores dimensiones, fin último del proyecto: la sistematización y la informatización codificada de todos y cada uno de los errores registrados en el corpus, para su análisis en trabajos de investigación especializados y para la preparación de materiales didácticos que puedan usarse en la enseñanza de español como lengua extranjera.

2.1. Recogida de materiales.

La creación y preparación del corpus se han efectuado con la ayuda de la dirección y el profesorado de los Cursos de Lengua y Cultura Españolas para Extranjeros de la Universidad de Alcalá, de los que hemos obtenido los materiales de producción escrita que lo conforman[6].

La recogida de materiales se ha realizado de forma periódica a lo largo de casi todo el año 2000. Con ello hemos pretendido obtener el mayor número posible de composiciones de estudiantes de los distintos niveles de aprendizaje, las cuales nos permitirán observar, además, la progresión de un mismo estudiante, al tener las producciones que ha escrito en distintos niveles de adquisición a lo largo de un trimestre o, incluso, un año, dependiendo de la duración de los cursos o del número de cursos en los que se matricule un alumno.

Aprovechando nuestros recursos, hemos trabajado con los cuatro niveles genéricos que se reconocen en nuestra Universidad: elemental, intermedio, avanzado y superior, aunque no descartamos la posibilidad de establecer subniveles en una etapa posterior del proyecto e incluir, por ejemplo, un quinto de perfeccionamiento. Además, dada la configuración de los Cursos de Lengua y Cultura Españolas para Extranjeros, nuestros informantes son de distinta caracterización social y de diversa procedencia lingüística, pudiendo controlar con ello la variabilidad condicionada por factores sociales y geográficos. Los datos personales de cada estudiante se recogen en una ficha que él mismo completa, al comienzo de cada nuevo curso, y que presenta el siguiente formato:

FICHA DE DATOS PERSONALES

- NOMBRE Y APELLIDOS:
- SEXO:
- EDAD:
- NACIONALIDAD:
- LENGUA(S) MATERNA(S) (Especificar manera de adquisición):
- 1ª LENGUA EXTRANJERA APRENDIDA (o segunda lengua):
- OTRAS LENGUAS QUE HABLA:
- ESTUDIOS REALIZADOS DE ESPAÑOL Y LUGAR DE
- REALIZACIÓN:
- NIVEL EN EL QUE SE ENCUENTRA:

[6]. Los estudiantes de español a los que se les pidieron las composiciones que forman el corpus han dado su autorización para utilizarlas.

Las producciones escritas con las que contamos son de dos tipos: una composición controlada semanal, realizada por cada uno de los estudiantes que han pasado por el centro como tarea para casa (con ayuda de materiales de apoyo), y una composición controlada mensual, realizada por cada uno de los estudiantes como tarea de clase (sin ayuda de materiales complementarios).

2.2. Almacenamiento del hábeas.

A medida que ha ido creciendo el corpus, se ha ido procediendo a su almacenamiento y etiquetado informático, con el fin de que, en un futuro inmediato, sea de fácil acceso y uso para todo aquel interesado en el estudio de algún aspecto relacionado con la enseñanza o el aprendizaje del español como lengua extranjera.

El material recogido se ha almacenado electrónicamente en dos fases y en sendos formatos. En primer lugar, como documentos de Word 97, el procesador de textos de Microsoft© que se ha empleado en la digitalización de las composiciones. Además, se ha creado una base de datos, cuya estructura general y finalidad se describen más abajo.

En la primera fase, se han creado archivos independientes para cada estudiante que se distribuyen en una serie de carpetas ordenadas por niveles. De esta forma es relativamente fácil localizar y revisar los textos provenientes de un estudiante determinado. Todos los archivos tienen idéntica estructura, puesto que se han confeccionado a partir de una plantilla. Cada uno de ellos comienza con una cabecera, en la que se recogen los datos personales del estudiante, tras la cual van colocadas, de forma correlativa y numeradas, las composiciones que ha realizado el alumno o la alumna en un curso y nivel determinados[7]. El mecanografiado de cada composición es revisado, antes de concluir su primer almacenamiento, por dos personas diferentes al menos. Una vez terminada la primera fase de almacenamiento, se traslada, en una segunda fase, toda la información a la base de datos que se ha creado para este fin. Los campos que configuran dicha base son los siguientes:

CAMPOS

- APELLIDOS:
- NOMBRE[8]:
- SEXO:
- EDAD:
- NACIONALIDAD:
- LENGUA MATERNA 1:
- LENGUA MATERNA 2:
- LENGUA APRENDIDA[9]:
- OTRAS LENGUAS:

[7]. De esta manera podemos comprobar el progreso de un estudiante en un mismo nivel.
[8]. En la base de datos definitiva, los apellidos y el nombre del informante son sustituidos por un código numérico para salvaguardar la confidencialidad de los datos personales.
[9]. Primera lengua aprendida como lengua extranjera.

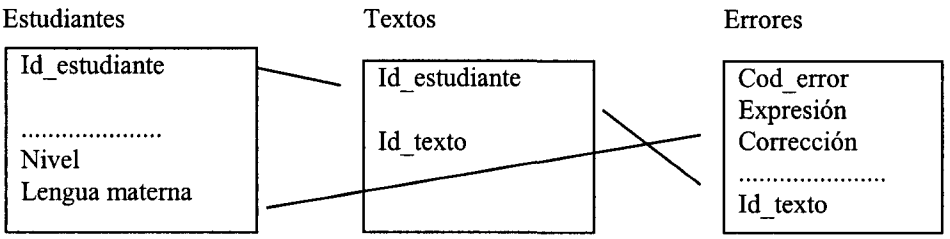
- ESTUDIOS REALIZADOS DE ESPAÑOL Y LUGAR DE REALIZACIÓN:
- NIVEL EN EL QUE SE ENCUENTRA:
- TIPO DE COMPOSICIÓN (con ayuda de materiales de apoyo - sin ayuda de materiales de apoyo):
- COMPOSICIÓN:

Esta base de datos se ha creado, inicialmente, empleando Access 97 de Microsoft©, pero está prevista la migración a SQL Server de Microsoft© o DB2 de IBM©, si el tamaño final del corpus lo hiciese aconsejable. La organización de los datos en una serie de tablas supone notables ventajas con respecto a la estructura plana de los ficheros de texto: facilita el proceso de anotación de errores, objetivo primordial del proyecto, permite filtrar y tabular los datos conforme a diferentes parámetros, así como efectuar cálculos y estadísticas, y posibilita la exportación automática de los datos a diversos formatos (HTML, SGML, TEI, RTF), con miras a la publicación electrónica o en papel de datos y resultados.

La base de datos consta, en principio, de tres tablas principales: una *(Estudiantes)* en la que se almacenan los datos de los estudiantes, otra *(Textos)* en la que se almacenan los textos y una tercera *(Errores)* en la que se recogen, debidamente clasificados, los errores que se anotan en aquellos. Además de estas tres tablas fundamentales, se define otra serie de tablas en las que se organizan jerárquicamente, con estructura de tesauro, los códigos de errores. Estas tablas serán el soporte de la herramienta de anotación que en estos momentos estamos preparando.

La tabla *Estudiantes* contiene todos los datos relativos a cada uno de los estudiantes que han servido de muestra y que ya se recogen en la cabecera del documento de Word correspondiente, más un código identificador único de cada estudiante. La tabla *Textos* recoge el identificador del estudiante, otro identificador con el número de la composición, una clave que indica si el texto está marcado o no y el propio texto. La tabla *Errores* contiene el código de error, la expresión incorrecta, la expresión correcta (para poder generar, en su caso, ejemplos y ejercicios de forma automática), el identificador de la composición y punteros que localizan esta expresión en el texto correspondiente.

De esta forma, se economiza el mecanografiado de datos y es fácil relacionar cada error con todos los datos relativos al mismo, dentro de la propia tabla *Errores* o mediante relaciones con las otras tablas a través de los identificadores. Por ejemplo, podemos listar o contar cuántos estudiantes con una lengua materna dada han cometido un error determinado, estableciendo una relación virtual entre ambos datos a través de las relaciones entre Id_texto e Id_estudiante.



La base de datos, en la que cada composición constituye una entrada diferente, aunque varias de ellas pertenezcan a un mismo estudiante en un mismo nivel de aprendizaje, nos permite clasificar el corpus atendiendo a variables lingüísticas, geográficas y sociales. Así, el corpus se puede organizar, dependiendo del tipo de análisis que se desee realizar sobre él, a partir: 1º del nivel de aprendizaje en el que se encuentran los estudiantes (elemental, intermedio, avanzado y superior), 2º del sexo (hombre o mujer), 3º de la edad (se han establecido los siguientes grupos de edad: 16-20; 21-25; 26-30; 31-35; 36-40; 41-45; 46-50; 51-55 y más de 55), 4º del lugar de procedencia de los informantes (su lengua materna), 5º de la primera lengua aprendida como lengua extranjera, 6º del tiempo dedicado al estudio del español, 7º del lugar en el que se han cursado los estudios y 8º del tipo de composición realizada (con materiales de apoyo o sin ellos).

2.3. El corpus en cifras.

En mayo de este año dimos por concluida la primera fase de recogida de materiales. Una vez eliminados los textos defectuosos, los resultados finales del corpus son los siguientes: el total de composiciones recogidas, almacenadas y etiquetadas es 1091, 1011 realizadas con materiales de apoyo y 80 sin ellos. 43 de ellas han sido realizadas por estudiantes de nivel elemental, 432 pertenecen a alumnos de nivel intermedio, 213 a estudiantes de nivel avanzado y 403 a estudiantes de nivel superior. Hasta el momento, han participado en nuestro corpus 321 estudiantes diferentes: 236 mujeres y 85 hombres, que tienen como lengua materna el japonés (58 estudiantes), el inglés (57), el alemán (36), el francés (29), el sueco (29), el italiano (18), el portugués (18), el japonés (22), el coreano (16), el chino (9), el turco (6), el húngaro (4), el polaco (3), el griego (3), el árabe (4), el ruso (2), el fang (2), el holandés (2), el flamenco (1), el danés (1), el finés (1), el esloveno (1), el pakistaní (1) y el checo (1); además hemos contado con 19 estudiantes bilingües (ewe/inglés, francés/flamenco, francés/inglés, francés/alemán, ruso/polaco, inglés/irlandés, inglés/africano, etc.).

3. ANÁLISIS DEL CORPUS

A partir de este momento, nuestro proyecto más inmediato es establecer una serie de marcas para identificar los errores que se manifiestan en la palabra entendida como signo: errores en el significante de la palabra, errores en el significado categorial de la palabra y errores en el significado léxico de la palabra, teniendo en cuenta, además, que las palabras se distribuyen en clases y se combinan para formar sintagmas, oraciones y textos. Las marcas en cuestión serán incluidas en las composiciones informatizadas, de forma que los distintos tipos de errores, cometidos por los estudiantes de E/LE, puedan ser rápidamente localizados en la base de datos.

El corpus así almacenado, etiquetado y marcado servirá, sin duda, para realizar una amplia gama de estudios; por ejemplo, la aplicación de métodos estadísticos al corpus permitirá establecer cuáles son los errores más frecuentes en el aprendizaje del español

como lengua extranjera en general o bien en función del nivel en que se encuentran los estudiantes o dependiendo de las distintas lenguas maternas habladas por ellos. Por otra parte, la marcación de los errores permitirá establecer las causas o el origen de los mismos en relación con los criterios que habitualmente se han establecido en el análisis de errores. Finalmente, los análisis obtenidos serán de gran utilidad para elaborar materiales didácticos para la enseñanza del español a extranjeros, en general, y para la enseñanza del español a hablantes de una lengua materna concreta.

4. REFERENCIAS BIBLIOGRÁFICAS

Álvarez Martínez, A. Á., A. Blanco Canales, M. L. Gómez Sacristán y N. Pérez de la Cruz (2000): *Sueña 1. Nivel Inicial*, Madrid, Anaya.

Aznar Juan, M. L. (1998): *Análisis de errores en el significado léxico de la producción escrita de estudiantes estadounidenses de español,* Memoria de investigación inédita para el *Máster en Enseñanza de Español como Lengua Extranjera*, Alcalá de Henares, Universidad de Alcalá.

Blanco Canales, A., M. C. Fernández López y M. J. Torrens Álvarez (2001): *Sueña 4. Nivel Superior*, Madrid, Anaya.

Cabrerizo Ruiz, M. A., M. L. Gómez Sacristán y A. Ruiz Martínez (2000): *Sueña 2. Nivel Medio*, Madrid, Anaya.

Damas Gil, C. (2000): *Los valores de los tiempos verbales de los modos indicativo y subjuntivo en español y en portugués. Análisis de errores en el significante y en los valores de los morfemas de tiempo de las formas verbales españolas,* Memoria de investigación inédita para el *Máster en Enseñanza de Español como Lengua Extranjera*, Alcalá de Henares, Universidad de Alcalá.

Duarte, C. A. (1997): *Análisis contrastivo de las preposiciones portuguesas y españolas. Análisis de errores en el uso de las preposiciones españolas por lusohablantes brasileños*, Memoria de investigación inédita para el *Máster en Enseñanza de Español como Lengua Extranjera*, Alcalá de Henares, Universidad de Alcalá.

Fuente, V. de la, I. Giraldo Silverio, F. Martín Martín, B. Sanz Sánchez Y M. J. Torrens Álvarez (2001): *Sueña 3. Nivel Avanzado*, Madrid, Grupo Anaya.

Gutiérrez Quintana, E. (2000): *Análisis de errores en la producción escrita de italianos aprendices de E/LE,* Memoria de investigación inédita para el *Máster en Enseñanza de Español como Lengua Extranjera,* Alcalá de Henares, Universidad de Alcalá.

Li, L. (1999): *Errores de concordancia en la producción escrita de estudiantes chinos*, Memoria de investigación inédita para el *Máster en Enseñanza de Español como Lengua Extranjera*, Alcalá de Henares, Universidad de Alcalá.

Lin, T.-J. (1998): *Análisis de errores en la expresión escrita de estudiantes adultos de español cuya lengua materna es el chino,* Memoria de investigación inédita para el *Máster en Enseñanza de Español como Lengua Extranjera,* Alcalá de Henares, Universidad de Alcalá.

Mansilla Clemente, F. (1999): *Funciones del SE en español y portugués de Brasil. Análisis contrastivo y análisis de errores,* Memoria de investigación inédita para el *Máster en Enseñanza de Español como Lengua Extranjera,* Alcalá de Henares, Universidad de Alcalá.

Morimoto, Y. e I. Penadés (2001): *Ejercicios de gramática española para hablantes de japonés. Nivel: inicial-intermedio-avanzado,* Madrid, Arco/Libros.

Penadés Martínez, I., coord. (1999): *Lingüística contrastiva y análisis de errores (español-portugués y español-chino)*, Madrid, Edinumen.

Sierra Martínez, F. (2000): "Algunos errores morfosintácticos en la expresión escrita del español como L2", en M. Franco Figueroa / C. Soler Cantos / J. de Cos Ruiz / M. Rivas Zancarrón / F. Ruiz Fernández (eds.): *Nuevas perspectivas en la enseñanza del español como lengua extranjera. Actas del X Congreso Internacional de ASELE (Cádiz, 22-25 de septiembre de 1999),* Cádiz, Servicio de Publicaciones de la Universidad de Cádiz, 663-671.

Tomazini, V. (1997): *Análisis de errores referidos a las categorías gramaticales en la producción escrita en español de alumnos lusohablantes brasileños*, Memoria de investigación inédita para el *Máster en Enseñanza de Español como Lengua Extranjera*, Alcalá de Henares, Universidad de Alcalá.