Pasado, presente y futuro de los corpus de aprendices de ELE. Una revisión bibliográfica

Eugenia Esperanza Núñez Nogueroles. Universidad de Granada

Recepción: 01/03/2019 | Aceptado: 30/04/2019

Correspondencia a través de **ORCID**: Eugenia Esperanza Núñez **0000-0003-0540-4242**

Citar: Núñez Nogueroles, EE (2019). Pasado, presente y futuro de los corpus de aprendices de ELE. Una revisión bibliográfica. *ReiDoCrea - Monográfico sobre Perspectivas transnacionales en la enseñanza de lenguas, 8*(3), 170-190.

Resumen: Este trabajo presenta una revisión bibliográfica en la que se examinan publicaciones referentes a los corpus de aprendices en general y a los de ELE en particular, destacando el papel fundamental de los avances informáticos en el desarrollo de este ámbito. Dado que la investigación en corpus de aprendices aborda el estudio de los errores, comenzaremos remontándonos a los orígenes de la Lingüística Contrastiva y sus tres modelos de análisis para, más adelante, poner de manifiesto que la ayuda proporcionada en la actualidad por los ordenadores ha solventado problemas detectados en los trabajos que implementaron estos modelos (en especial, el Análisis de Errores) durante el siglo XX. Nos centraremos, a continuación, en la investigación en corpus de aprendices: expondremos sus características principales y describiremos cinco de ellos. Exploraremos el Análisis de errores asistido por ordenador y el Análisis Contrastivo de Interlengua. Indagaremos en la relación entre la investigación basada en corpus de aprendices y la Adquisición de Segundas Lenguas, por un lado, y entre aquella y la Enseñanza de Lenguas Extranjeras, por otro. Haremos referencia a la producción pedagógica derivada del tipo de estudios que nos ocupan y, por último, incluiremos una sección final, a modo de conclusión.

Palabras clave: Lengua extranjera, Lengua materna

Past, Present and Future of Spanish Learner Corpora: a literature review

Abstract: This piece of research presents a literature review on learner corpora in general and Spanish learner corpora in particular, highlighting the fundamental role played by computer developments in the growth of this research area. Since many studies using learner corpora deal with students' errors, we will start by going back to the origins of Contrastive Linguistics and its three models of analysis; this examination will be useful later in this paper for emphasising that the assistance provided by computers nowadays has solved problems detected in pieces of research that implemented these models (mainly Error Analysis) in the 20th century. We will focus on learner corpus research: their principal features will be explained and five of these corpora will be described. We will explore Computer-aided Error Analysis (CEA) and Contrastive Interlanguage Analysis (CIA), as well as the relationship between learner corpus research and Second Language Acquisition, on the one hand, and the former and Foreign Language Teaching, on the other hand. We will also make reference to the pedagogical materials that have derived from the studies in hand. Last, a closing section will be included.

Keywords: Foreign language, Mother tongue.

Introducción

Este trabajo presenta una revisión de la bibliografía relativa a los corpus de aprendices en general y a los de español como Lengua Extranjera –ELE– en particular, exponiendo las características principales de cada uno de ellos y destacando el potencial que este tipo de recopilaciones textuales informatizadas poseen dentro del ámbito de la investigación en adquisición de segundas lenguas y en enseñanza de lenguas extranjeras. Las múltiples posibilidades que los avances tecnológicos han traído consigo facilitan la obtención de datos y el tratamiento de estos. Gracias al uso de programas informáticos especializados, en la actualidad pueden llevarse a cabo estudios caracterizados por una precisión y un rigor que distan mucho de los que encontramos en los trabajos relativos a las producciones de aprendices de una lengua extranjera publicados hace unas décadas.

Dado que el estudio de los *errores* presentes en los textos (orales o escritos) de los estudiantes constituye uno de los principales ámbitos abordados por la investigación en corpus de aprendices, es pertinente remontarse a los inicios de la Lingüística Contrastiva y sus tres modelos de análisis (Análisis Contrastivo, Análisis de Errores y Análisis de Interlengua) para, más adelante, poner de manifiesto que la ayuda proporcionada en la actualidad por los ordenadores ha solventado problemas detectados en los estudios que implementaron estos modelos (en especial, el Análisis de Errores) durante la segunda mitad del siglo XX. Así pues, este será nuestro punto de partida.

Posteriormente, nos centraremos en la investigación basada en corpus de aprendices. Tras una introducción al tema, describiremos cinco corpus que recogen producciones (orales o escritas) de estudiantes de español como Lengua Extranjera: CORANE, CEDEL2, CAES, CAELE y SPLLOC. Seguidamente, abordaremos dos tipos de análisis que hacen uso de corpus electrónicos: el Análisis de errores asistido por ordenador y el Análisis Contrastivo de Interlengua. Más adelante, indagaremos en las publicaciones que han profundizado en la relación entre la investigación basada en corpus de aprendices y la Adquisición de Segundas Lenguas, por un lado, y entre aquella y la Enseñanza de Lenguas Extranjeras, por otro.

Proseguiremos haciendo referencia a la producción pedagógica derivada del tipo de estudios que nos ocupan. En concreto, expondremos las principales características de varios manuales de ELE dirigidos a estudiantes que poseen el inglés como lengua materna.

Cerraremos el presente artículo con una sección final, a modo de conclusión, en la que situaremos la investigación en corpus de aprendices en el momento actual y destacaremos sus perspectivas futuras.

Lingüística Contrastiva

La Lingüística Contrastiva es una "subdisciplina de la Lingüística Aplicada, que surge a mediados de los años cuarenta ligada al estructuralismo de signo conductista" (Santos Gargallo, 1993, p. 25).[1] Su propósito es contrastar, desde un punto de vista descriptivo y sincrónico, dos o más lenguas (normalmente la lengua materna de unos estudiantes y la lengua meta que estos se encuentran aprendiendo) para, de este modo, determinar las similitudes y diferencias estructurales que caracterizan a esos sistemas lingüísticos objeto de estudio.

Cronológicamente, podemos distinguir tres modelos de análisis dentro del ámbito de la Lingüística Contrastiva: el Análisis Contrastivo (AC), el Análisis de Errores (AE) y el Análisis de la Interlengua (AI). Las críticas al primero dieron lugar al surgimiento del segundo, y lo mismo ocurrió entre este y el tercero. A continuación describiremos cada uno de ellos, para finalmente cerrar esta sección llevando a cabo una revisión de la evolución del concepto de *error* a lo largo del tiempo.

Análisis Contrastivo (AC)

Este modelo propone utilizar los resultados obtenidos tras el contraste de dos lenguas dadas (a saber, las diferencias y semejanzas estructurales existentes entre ellas) para predecir las dificultades que encontrarán los estudiantes nativos de uno de estos idiomas cuando se hallen inmersos en el proceso de aprendizaje del otro.

[1] La obra *Análisis Contrastivo, Análisis de Errores e Interlengua en el marco de la Lingüística Contrastiva*, publicada por Isabel Santos Gargallo en 1993, constituye un estudio completo y exhaustivo que tomaremos como punto de referencia para la presente sección de este artículo.

Como leemos en Santos Gargallo (1993), el AC tiene su origen en los trabajos llevados a cabo por C. Fries (1945) y R. Lado (1957) en la Universidad de Michigan. Basándose en el estructuralismo lingüístico[2] de L. Bloomfield y en el conductismo psicológico (*behaviourism*, paradigma estímulo – respuesta), estos autores defienden que "el aprendizaje de una segunda lengua[3] es la formación de un hábito" (Santos Gargallo, 1993, p. 35) y presentan la hipótesis del AC: "Un hábito viejo (el de la lengua nativa) facilita la formación de nuevos hábitos (los de la lengua meta) dependiendo de las similitudes y diferencias entre los viejos hábitos y los nuevos" (1993, p. 35). Así pues, si se trata de dos lenguas entre las que abundan las semejanzas, tendrán lugar *transferencias positivas*[4] desde la lengua materna hacia la lengua meta, mientras que si el número de diferencias es elevado se producirán muchas *transferencias negativas* o *interferencias* durante el proceso de aprendizaje del idioma extranjero (Vez Jeremías, 2004).

Esas interferencias (fonéticas, morfológicas, sintácticas y léxicas) ocurren cuando el aprendiz utiliza en la lengua meta un rasgo de su idioma nativo y se da la circunstancia de que no existe correspondencia en cuanto a ese elemento[5] en las dos lenguas. Para los defensores del AC, esta es la principal y casi única causa de los errores que se cometen tanto en la recepción como en la producción de una lengua extranjera que se está aprendiendo. Basándose en esta idea, afirman que, si efectuamos un Análisis Contrastivo podremos descubrir, *a priori*, cuáles serán las áreas conflictivas a las que deberán hacer frente los estudiantes, y seremos capaces de adelantarnos a las situaciones problemáticas y prevenir los errores.

Los resultados obtenidos en diversos estudios empíricos dieron lugar a que, durante los años 60, el AC recibiera numerosas críticas, entre las que destaca el hecho de que no todos los errores se deben, como el AC propugnaba, a la interferencia interlingüística. Tras estas críticas, Wardhaugh (1970) reformuló los planteamientos iniciales del AC y distinguió dos versiones dentro de este modelo de análisis:

> La versión fuerte que pretendía diagnosticar y predecir los errores como resultado de la comparación de la LM y la lengua meta sin recurrir necesariamente a las producciones reales de los hablantes (errores de interferencia); y otra, la versión débil, que intentaba describir las dificultades y explicar los errores cometidos por los hablantes (y no, como la versión fuerte, predecir su aparición) (Zimny, 2016, p. 37).

Análisis de Errores (AE)

A finales de los años 60, S. Pit Corder (1967) sienta las bases de un modelo de análisis que supone un avance en el ámbito de la Lingüística Contrastiva. El autor se inspira en los postulados de Noam Chomsky, quien dos años antes[6] había cuestionado tanto el conductismo psicológico sobre el que se cimentó el Análisis Contrastivo como la teoría

[2] No obstante, el AC fue evolucionando a lo largo del tiempo y se produjeron diversos cambios en cuanto a la teoría lingüística utilizada: estructuralista, generativo-transformacional, psicolingüística, mixta –estructuralista y generativista– y, en muchas ocasiones, ecléctica –consistente en seleccionar lo que se considera mejor de cada modelo– (Santos Gargallo, 1993).

[3] La diferencia entre *segunda lengua* (L2) y *lengua extranjera* (LE) radica en el contexto extralingüístico que rodea al aprendiz. Como se indica en el *Diccionario de términos clave de ELE*, "cuando la L[engua] M[eta] se aprende en un país donde no es ni oficial ni autóctona, se considera una LE (…). Cuando la L[engua] M[eta] se aprende en un país donde coexiste como oficial y/o autóctona con otra(s) lengua(s), se considera una L2 (…). Otra situación cada vez más frecuente de L2 es la que se produce cuando los emigrantes con una L1 común llegan a constituir una comunidad de habla relevante en el país de acogida".

[4] Transferencia positiva: "es el fenómeno resultante de emplear con éxito comunicativo elementos propios de una lengua (mayormente, la L1) en otra lengua" (Centro Virtual Cervantes*, Diccionario de términos clave de ELE*).

[5] En palabras de López Salinas (2001, p. 102), la estructura de que se trate no tiene un "comportamiento lingüístico similar en ambos sistemas".

[6] *Aspects of the Theory of Syntax* (1965).

de adquisición de lenguas de Skinner. Chomsky propugna la existencia de una Gramática Universal[7] y defiende que "la adquisición de lenguas no es el resultado de una formación de hábitos, sino de una generación de reglas" (Belda Torrijos, 2015, p. 93). El AE adopta la lingüística generativo-transformacional y el cognitivismo como modelos teóricos (Santos Gargallo, 1993).

El AE se puede definir como una "técnica de observación, identificación, análisis, clasificación e interpretación de las producciones idiosincrásicas de los hablantes no nativos, en cualquier situación espontánea o controlada de respuesta lingüística" (Baralo Ottonello, 2009, p. 27). Este modelo estudia, de manera sistemática, los errores producidos por los estudiantes de una lengua extranjera.

Por tanto, "la diferencia entre el modelo de análisis contrastivo y el de análisis de errores radica en que en este modelo no se parte de la contrastación de los pares de lenguas, la propia y la estudiada, sino de las producciones reales de los aprendices en su contexto" (Belda Torrijos, 2015, p. 95).[8] Así pues, el AC se realiza *a priori* y se caracteriza por ser abstracto, mientras que el AE se lleva a cabo *a posteriori* y se basa en producciones concretas.

Los estudios empíricos desarrollados dentro del marco del AE han contradicho creencias defendidas por el AC. Por ejemplo, se ha puesto de manifiesto que el grado de divergencia entre la lengua materna y la lengua meta no determina necesariamente la probabilidad de error. Asimismo, estos estudios han señalado la conveniencia de que el investigador pudiera corroborar su interpretación de los errores mediante información proporcionada por el alumnado –es decir, preguntándoles a los estudiantes qué les indujo a producir una determinada unidad/secuencia lingüística– (Santos Gargallo, 1993).

Es interesante mencionar el giro que el propio Corder da al Análisis de Errores a principios de los años 80. Hasta ese momento, los trabajos que implementaban un AE centraban su atención en la competencia gramatical del estudiante y clasificaban los errores según la categoría gramatical a la que estos pertenecían. Con el nuevo planteamiento introducido por Corder en 1981, el AE va a abarcar también la evaluación del resto de las categorías que forman la competencia comunicativa, "incluyendo nuevos criterios de índole pragmática y semántica" (Santos Gargallo, 1993, p. 88). Además, tras haber recibido el AE críticas[9] por reflejar únicamente lo que el aprendiz hace mal en lugar de adoptar una perspectiva global, el autor incorpora otro aspecto novedoso que tendrá especial relevancia en el Análisis de la Interlengua:[10] es necesario estudiar no solo las producciones erróneas sino también las correctas, ya que esta es la única manera de obtener una visión completa de la competencia del estudiante en un momento determinado.

Análisis de Interlengua (AI)

El término *interlengua* (IL) fue acuñado por L. Selinker en 1969 y reelaborado por el mismo autor en su artículo "Interlanguage" (1972). En este trabajo, Selinker define la

[7] Gramática Universal (GU): "conjunto de principios, reglas y condiciones que comparten todas las lenguas. Este concepto constituye el núcleo de la teoría de la gramática generativo-transformacional, con la que N. Chomsky propuso explicar el proceso de adquisición y uso de la lengua. Según esta teoría, todos los seres humanos adquieren de forma natural una lengua cualquiera porque disponen de una gramática universal" (*Diccionario de términos clave de ELE*).

[8] En palabras de Sánchez Rufat y Jiménez Calderón (2013, p. 185), "[p]or primera vez, la lengua del aprendiente era el principal foco de atención de los investigadores, y no ya la lengua materna (L1) y la lengua meta (L2) como se venía haciendo en el marco del *análisis contrastivo*" (la cursiva es original).

[9] Otros juicios negativos que se le han aplicado al modelo de AE hacen referencia a la imposibilidad de encontrar, en ocasiones, un único origen para cada error y al "problema conceptual entre explicación y descripción de los errores" (de Alba Quiñones, 2009, p. 12).

[10] *Véase* el apartado *Análisis de Interlengua (AI)* del presente trabajo.

interlengua como "a separate linguistic system based on the observable output which results from a learner's attempted production of a T[arget] L[anguage] norm"[11] (Selinker, 1972, p. 214). Para lanzar la hipótesis de que este sistema lingüístico independiente existe realmente, el autor argumenta que el conjunto de emisiones lingüísticas de la mayoría de los estudiantes de una segunda lengua no es idéntico al hipotético conjunto correspondiente de emisiones lingüísticas que produciría un hablante nativo de esa lengua meta para expresar el mismo significado.

A partir de esta afirmación, Selinker (1972) plantea que, para estudiar los procesos psicolingüísticos que subyacen al comportamiento de la IL, los únicos datos observables que podemos considerar relevantes son los siguientes:[12]

- Las emisiones en la lengua materna del aprendiz producidas por el propio aprendiz
- Las emisiones en la interlengua producidas por el aprendiz
- Las emisiones en la lengua meta producidas por hablantes nativos de la lengua meta

La IL "presenta elementos de la lengua materna, otros de la lengua meta y algunos exclusivamente idiosincrásicos" (Santos Gargallo, 2004, p. 393). En cuanto a los procesos psicolingüísticos a los que acabamos de hacer referencia, Selinker indica que los cinco que comentaremos a continuación son centrales en el aprendizaje de segundas lenguas (*second-language learning*):[13]

1. La transferencia lingüística: Desde la lengua materna a la lengua meta.
2. La transferencia de instrucción: "A partir de la explicación o la enseñanza formal, el aprendiz intenta aplicar los nuevos conocimientos sin conocer bien los límites de su uso" (Varón López, 2008, p. 110).
3. Las estrategias de aprendizaje de segundas lenguas: Se refieren a la simplificación del sistema de reglas y categorías de la lengua meta.
4. Las estrategias de comunicación en segundas lenguas: Existen dos grandes grupos de estrategias que son empleadas con el objetivo de solucionar problemas de comunicación: las estrategias de reducción y las estrategias de expansión o de logro de éxito.[14]
5. La sobregeneralización del material lingüístico de la lengua meta: "Se aplica una regla de la L2 a todas las situaciones parecidas, traspasando los límites de su uso correcto" (Varón López, 2008, p. 109).

Santos Gargallo (1993, 2004) señala que la interlengua ha recibido también otros nombres, como *dialecto idiosincrásico* o *competencia transitoria* (Corder, 1967, 1971) y *sistema aproximativo* (Nemser, 1971) –dado que se va aproximando, cada vez más, a la lengua meta–. Esta autora hace asimismo referencia a diversos investigadores que han sugerido la idea de *continuum* en la IL, describiéndola como un proceso dinámico e inestable que va evolucionando y atraviesa etapas sucesivas en las que su complejidad aumenta a medida que el aprendiz adquiere nuevos conocimientos:

[11] Un sistema lingüístico separado basado en el comportamiento lingüístico observable que resulta de los intentos de producción de la norma de la lengua meta que realiza el aprendiz (traducción propia).

[12] Solo se tendrán en cuenta las situaciones de actuación dotadas de significado (*meaningful performance situations*), i.e. aquellas situaciones en las que un adulto –definido como una persona mayor de 12 años– intenta expresar significados –cuyo conocimiento puede ya poseer– en una lengua en cuyo proceso de aprendizaje se encuentra inmerso. Esto excluye, por tanto, producciones como las que resultan de la repetición de *drills*.

[13] "first, *language transfer;* second, *transfer-of training;* third*, strategies of second-language learning;* fourth*, strategies of second-language communication;* and fifth*, overgeneralization of TL linguistic material*" (Selinker, 1972, p. 215).

[14] Para una descripción detallada de estas estrategias, *véase* Santos Gargallo (1993, pp. 141-148).

Esta *lengua del aprendiz* (Zimny, 2016) se caracteriza no solo por las producciones desviadas [15] (con respecto a la norma de la lengua meta) sino también por las correctas. [16] Describiendo ese "sistema no-nativo" arrojaremos luz sobre "la especificidad y naturaleza de la idiosincrasia de la IL" (López Salinas, 2001, p. 106).

Evolución del concepto de error

Para delimitar qué entendemos por *error* haremos referencia a Torijano (2006), quien distingue entre *error*, *equivocación* y *lapsus*. Con relación al primero, indica que "[ú]nicamente debería llamarse "error" a la 'desviación sistemática, producida por el hecho de que un estudiante todavía no haya aprendido algo y, consecuentemente, lo expresa mal'" (p. 154). Con respecto al segundo, considera que se trata de una alternancia, una "*desviación incoherente*, muestra de inseguridad en las producciones, debida a causas no siempre conocidas (…): unas veces el estudiante lo dice bien, pero otras se equivoca y usa la forma incorrecta" (p. 156; la cursiva es original). Finalmente, Torijano define el *lapsus* como una "desviación debida a una falta de concentración, a un fallo de memoria, al cansancio, etc." que "se produce como consecuencia de la propia actividad cerebral" (p. 156) y es susceptible de ser cometido por cualquier usuario de una lengua, no solo por los aprendices de esta.

A lo largo de los años a los que nos hemos referido en los subapartados anteriores, el error ha sido considerado de diferentes formas. En líneas generales, podemos afirmar que se ha pasado de una actitud negativa ante las incorrecciones, en un primer momento, a una visión positiva del error en las etapas más recientes.

Los defensores del Análisis Contrastivo tenían "el objetivo de evitarlos, en la creencia conductista de que así evitarían la formación de malos hábitos en la lengua meta" (Baralo Ottonello, 2009, p. 28). Desde esta perspectiva, el error es algo negativo, un "síntoma de fracaso" (Torijano, 2006, p. 149).

Posteriormente, el Análisis de Errores revaloriza el error, dotándolo de una carga positiva por ser indicador de que el proceso de aprendizaje está teniendo lugar. Se consideran inevitables y necesarios para la evolución del estudiante en su conocimiento de la lengua meta. Dentro del marco del AE, diversos autores han propuesto diferentes taxonomías de errores con el objetivo de poder clasificarlos de manera sistemática. En su tesis doctoral, Zimny (2016, pp. 51-55) presenta las tipologías más comunes según distintos criterios:[17]

1. Criterio gramatical (error en las estructuras gramaticales)
   - fonológicos
   - ortográficos
   - léxico-semánticos
   - morfosintácticos
   - pragmáticos

[15] Tomamos el término *desviada* de la definición de *sistema aproximativo* proporcionada por W. Nemser (1971, p. 116): "sistema lingüístico desviado empleado por el estudiante al intentar usar la lengua meta". Es interesante el comentario que, a propósito de esta definición, aporta Santos Gargallo (1993, p. 127): "la palabra «desviado» [es] muy significativa y [está] ausente en Corder y Selinker, para quienes la lengua del estudiante es, ante todo, un sistema lingüístico autónomo con su propia gramática, y un sistema correcto en su propia idiosincrasia". Esta última postura considera, por tanto, que las producciones que serían erróneas en la lengua meta son esperables y normales en la interlengua.

[16] Así pues, un Análisis de Interlengua atenderá a ambos tipos de producciones, como se observa en trabajos de autores como Ruiz Martín (2006).

[17] Aquí sintetizamos en un esquema la información presentada por Zimny (2016). Para una descripción detallada de cada tipología, ilustrada con ejemplos, *véase* la obra de esta autora.

2. Criterio lingüístico (también llamado descriptivo o de estrategias)
   - de adición
   - de omisión
   - de selección falsa
   - de colocación falsa
   - de yuxtaposición
3. Criterio etiológico (relativo a las causas de los errores)
   - intralinguales
   - interlinguales
   - de simplificación
4. Criterio comunicativo (relacionado con la transmisión del mensaje)
   - de ambigüedad
   - irritantes
   - estigmatizantes
   - globales frente a locales
   - errores de falta de pertinencia con respecto al contexto
   - errores de diversión[18]
5. Criterio pedagógico
   - inducidos frente a creativos
   - transitorios frente a permanentes
   - fosilizados frente a fosilizables
   - individuales frente a colectivos
   - residuales frente a actuales
   - de producción oral frente a producción escrita
6. Criterio pragmático[19]
   - de pertinencia o discursivos
7. Criterio cultural

Entre los tipos de error enumerados, aparecen dos que aluden a un fenómeno que merece la pena destacar: la *fosilización*.[20] Se trata del mecanismo por el cual "el HNN[21] tiende a conservar en su IL formas, reglas y subsistemas erróneos, de manera recurrente y en estadios del aprendizaje en que estos esquemas deberían estar superados" (Santos Gargallo, 2004, p. 394). Como vemos dentro del "criterio pedagógico", existen los errores fosilizables y los fosilizados. A continuación reproducimos la definición que de cada uno de ellos proporciona Fernández López (1995):

> [Los errores fosilizables] son aquellos que se repiten en fases sucesivas y que ofrecen una mayor resistencia, ya sea por la complejidad misma de la estructura, por un problema de interferencia o por cualquier otra clase de contaminación (Fernández López, 1995, p. 212).
>
> [Los errores fosilizados] son aquellos que se han fijado y no evolucionan ya; son los típicos errores de los grupos de población extranjera que se integran poco en la sociedad y que llegan a constituir una variedad de la norma estándar, o aquellos otros menos serios, que pueden aparecer

[18] Quizás *diversión* no sea la palabra más adecuada en español para referirse a un tipo de error comunicativo. Debido a que Zimny no aporta ningún ejemplo en este punto (únicamente indica que los errores de diversión "distraen la comunicación"), no es posible proponer con certeza una alternativa terminológica acertada. Sin embargo, es factible que la elección del vocablo *diversión* por parte de la autora pueda estar inducida por una confusión con la voz inglesa *diversion*, que significa 'desviación', 'desvío' (en cuanto al tráfico), 'distracción'.

[19] La autora no clarifica el motivo por el que incluye los errores pragmáticos tanto dentro de los gramaticales como en una categoría aparte. Quizá los haya encontrado de las dos maneras en distintas tipologías elaboradas por diferentes autores. No obstante, una nota aclaratoria habría resultado útil en este punto.

[20] Para profundizar en el tema de la fosilización, *véase* Bustos Gisbert y Sánchez Iglesias (2006).

[21] Hablante No Nativo.

> involuntariamente en situaciones especiales de cansancio, nerviosismo y que el mismo autor enmienda. Estos errores no parecen corregibles y la mejor postura ante ellos es la tolerancia (Fernández López, 1995, p. 213).

Con el objetivo de subsanar los problemas detectados con respecto a la diferenciación entre descripción y explicación de los errores en el modelo del AE (*véase* nota 9), Alexopoulou (2006) presenta un interesante artículo en el que delimita con precisión estos dos ámbitos. La autora asigna sendos criterios de clasificación de errores a dos de las tres etapas principales de la metodología del AE establecidas por Corder (1971) –1. identificación; 2. descripción; 3. explicación del error–:

- Etapa de la descripción de los errores (hace referencia al *producto*): criterio descriptivo
- Etapa de la explicación de los errores (hace referencia al *proceso*): criterio etiológico

En cuanto al primer criterio, Corder (1973) defiende que este se puede llevar a cabo en, al menos, dos niveles:

1. Descripción de los errores con relación a "la diferencia física entre el enunciado del estudiante y la versión reconstruida" (Alexopoulou, 2006, p. 20). Se trata de un tipo de descripción superficial que constituye el criterio descriptivo propiamente dicho. Dentro de este nivel se distinguen cuatro categorías:
   a.) "Omisión de algún elemento obligatorio"
   b.) "Adición de algún elemento innecesario o incorrecto"
   c.) "Selección de un elemento incorrecto"
   d.) "Mal ordenamiento de los elementos" (Alexopoulou, 2006, p. 20).
2. Descripción más profunda atendiendo a los distintos niveles del sistema lingüístico. Es lo que se conoce como criterio lingüístico:
   a.) Fonético-fonológico-ortográfico
   b.) Morfológico
   c.) Sintáctico
   d.) Léxico-semántico
   e.) Discursivo
   f.) Pragmático

En relación con el segundo criterio, el referido al descubrimiento de las causas que motivan el error, Alexopoulou (2006) advierte de la complejidad que ello supone, dado que hay errores ambiguos que poseen más de una razón de ser. Tras llevar a cabo una revisión de tipologías de causas de errores propuestas por distintos autores, Alexopoulou plantea establecer dos grandes grupos:

- Errores interlinguales: motivados por la interferencia de la lengua materna o de otra que el estudiante ha aprendido con anterioridad.
- Errores intralinguales: los que se deben a las complejidades internas de la lengua meta.

Así pues, la autora diseña un procedimiento de estudio según el cual debemos, en primer lugar, clasificar los errores atendiendo a un criterio descriptivo (y lingüístico) y, en segundo lugar, apoyándonos en esa fase de descripción, aplicar un criterio etiológico que nos permita determinar la causa (interlingüística o intralingüística) que ha motivado su aparición.

Investigación basada en Corpus de Aprendices

El *Diccionario de términos clave de ELE* define la Lingüística de Corpus como

> una rama de la lingüística que basa sus investigaciones en datos obtenidos a partir de corpus, esto es, muestras reales de uso de la lengua. En rigor, el término no define una disciplina lingüística (…), sino

> un enfoque metodológico que es posible adoptar desde disciplinas diversas.

Esta metodología nos proporciona herramientas para explotar los *corpus*, conjuntos de textos auténticos y en formato electrónico (incluyendo las transcripciones de datos orales) que están muestreados para ser representativos de una lengua o una variedad lingüística particular (McEnery, Xiao y Tono, 2006).

En la presente revisión bibliográfica nos centraremos en un tipo de corpus en concreto: los corpus de aprendices.

Los corpus de aprendices

Se entiende por corpus de aprendices (*learner corpora*) aquellas colecciones electrónicas de datos naturales o quasi-naturales producidos por estudiantes de lenguas segundas o extranjeras y recopilados de acuerdo con unos criterios de diseño explícitos (Granger, Gilquin y Meunier, 2015).

Los corpus de aprendices informatizados –CAI–[22] aparecieron a finales de los años 80 – principios de los 90 del siglo pasado (Ballier, Díaz-Negrillo y Thompson, 2013), pero fue a partir de 2002, con la publicación del *International Corpus of Learner English* (ICLE), cuando se empezaron a explotar estos corpus a gran escala. El ICLE, desarrollado por Sylviane Granger y su equipo en el Centre for English Corpus Linguistics (CECL) de la Universidad Católica de Lovaina, constaba (en su primera versión) de 2,5 millones de palabras y estaba compuesto por ensayos argumentativos producidos por estudiantes universitarios de inglés que poseían distintas lenguas maternas: español, italiano, francés, ruso, etc. (Lozano Pozo y Mendikoetxea, 2013).[23]

Este primer gran corpus de aprendices de inglés como lengua extranjera ha dado lugar a la creación de otros corpus de producciones de estudiantes (de diferentes L1) en esa misma lengua meta, como el *Longman Learner Corpus* (LLC) y el *Cambridge Learner Corpus* (CLC). En España se han compilado corpus de estudiantes nativos de español: en la Universidad Autónoma de Madrid se ha recopilado el *Written Corpus of Learner English* (WriCLE) y en la Universidad de Santiago se ha elaborado el *Santiago University Learner of English Corpus* (SULEC) (Lozano Pozo y Mendikoetxea, 2013).

Además de estos proyectos que han centrado su atención en el inglés como lengua meta, se han desarrollado corpus de aprendices destinados a estudiar otros idiomas como lenguas extranjeras. Debido a las limitaciones de espacio que constriñen toda publicación, en este trabajo nos centraremos en la lengua española. Así pues, nos disponemos ahora a presentar una selección de los principales corpus de estudiantes de ELE, tanto escritos como orales, recopilados hasta la fecha. Pero antes, merece la pena que nos detengamos en comentar someramente los criterios más relevantes que se deben tener en cuenta a la hora de diseñar o trabajar con un corpus (de cualquier tipo, entre ellos los de aprendientes que son los que nos ocupan), ya que esto nos ayudará a caracterizar mejor los corpus que abordaremos a continuación. Para ello, nos basaremos en la clasificación propuesta por Sánchez Rufat (2015):

1) Uso de corpus anotado *versus* no anotado

Un corpus anotado es aquel en el que se ha codificado, mediante el uso de etiquetas, información relativa a las palabras o secuencias de palabras que lo componen. La anotación puede referirse a la clase de palabra (sustantivo, adjetivo, verbo…), al tipo de error (de adición, de simplificación, global, local…) o a cualquier otra característica que

[22] Del inglés "computer learner corpus" (Granger, 2002, p. 3) o "computerised learner corpora" (Ballier, Díaz-Negrillo y Thompson, 2013, p. 3).

[23] "Para las comparaciones de interlengua vs. lengua nativa, se creó un corpus equivalente de inglés nativo *Louvain Corpus of Native English Essays* (LOCNESS)" (Lozano Pozo, 2009, p. 198).

queramos señalar para después poder llevar a cabo distintos análisis mediante herramientas informáticas de recuperación de datos.

2) Uso de un corpus longitudinal *versus* transversal

En el ámbito de los corpus de aprendices, a los corpus diacrónicos se les conoce como "longitudinales". Estos permiten analizar la evolución de la interlengua de un mismo grupo de estudiantes a lo largo del tiempo. Debido a la dificultad que entraña su compilación, una alternativa es la recopilación de producciones de aprendices de distintas edades o niveles (inicial, intermedio y avanzado) para formar un corpus cuya estructura se asemeje a la de los longitudinales. Por ello, reciben el nombre de "quasi-longitudinales". Por su parte, los corpus de aprendices sincrónicos son también llamados "transversales" (*cross-sectional*) y se usan para estudiar los textos (orales o escritos) producidos por un grupo determinado de aprendientes en un momento concreto.[24]

3) Análisis de corpus cuantitativo y cualitativo

El hecho de que los corpus contengan una ingente cantidad de palabras permite llevar a cabo análisis de tipo cuantitativo relativos a diversos aspectos. Para obtener resultados relacionados con, por ejemplo, la frecuencia de aparición de un fenómeno en la lengua, recurriremos a análisis estadísticos. Sin embargo, la recuperación de datos mediante búsquedas informatizadas no debería ser la única fuente de resultados cuando trabajamos con corpus de aprendices. Es necesario un análisis cualitativo posterior para poder llegar a conclusiones sólidas.

Tras abordar estos tres puntos referentes a la caracterización de los corpus y su estudio, presentamos a continuación una descripción de los principales corpus de aprendices de Español como Lengua Extranjera que se han compilado a lo largo de los últimos años.

CORANE (Corpus para el Análisis de Errores de Aprendices de E/LE)

Surge CORANE para dar respuesta a la necesidad de contar con un corpus de aprendices de ELE de fácil acceso y de grandes dimensiones que contuviera una recopilación de textos sistemática y organizada. Hasta ese momento, distintos investigadores habían compilado grupos reducidos de producciones de estudiantes que les resultaran útiles para cumplir los objetivos de sus trabajos. Ante esta situación, Cestero Mancera, Penadés Martínez y su equipo (Universidad de Alcalá) decidieron poner en marcha un proyecto que se proponía crear un corpus de materiales escritos, en una primera fase, e informatizarlo, almacenarlo, etiquetarlo y clasificar los textos según distintos parámetros, en una segunda etapa. Por último, se pretendía "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" (Cestero Mancera et al., 2002, p. 529).

La recogida de datos tuvo lugar en el año 2000 en los Cursos de Lengua y Cultura Españolas para Extranjeros de la Universidad de Alcalá. Participaron 321 estudiantes y se obtuvo un total de 1091 composiciones, las cuales se clasificaron de acuerdo con las siguientes variables: nivel del estudiante (elemental, intermedio, avanzado y superior); sexo (hombre o mujer); edad (se establecieron varias franjas); lugar de procedencia y lengua materna del informante; primera lengua aprendida como lengua extranjera; tiempo dedicado al estudio del español; lugar donde se han cursado dichos estudios; tipo de composición elaborada[25] (con o sin materiales de apoyo).

Este corpus de aprendices se publicó en formato CD-ROM en 2009.

---

[24] *Véase* también Granger (2009).

[25] "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)" (Cestero Mancera et al., 2002, p. 530).

CEDEL2 (Corpus Escrito del Español como L2)

Se trata de un corpus que se crea en el marco del proyecto *Word Order in Second Language Acquisition Corpora* (WOSLAC) –dirigido por Amaya Mendikoetxea (Universidad Autónoma de Madrid)–. En su origen contemplaba únicamente la inclusión de textos producidos por estudiantes cuya lengua materna era el inglés, pero posteriormente se amplió y en la actualidad alberga también datos de nativos de griego. Además, con el propósito de permitir estudios comparativos entre la interlengua de estudiantes de español y la lengua materna de los nativos de este idioma, se ha creado un subcorpus de hablantes nativos de español cuyos participantes siguen las mismas pautas que los aprendices. Dicho subcorpus constituye, aproximadamente, un 25% del tamaño total de CEDEL2, mientras que el subcorpus de estudiantes comprende el restante 75% (nivel elemental: 25%; nivel intermedio: 25%; nivel avanzado: 25%).[26]

Para la elaboración de CEDEL2, Lozano Pozo (Universidad de Granada) y su equipo han seguido los diez principios propuestos por Sinclair (2005) para la creación de corpus, aplicándolos de la manera que se muestra a continuación:[27]

Principio 1: *Contenido del Corpus*. Se determina según criterios externos.

Principio 2: *Representatividad*

- Se les proporciona a los estudiantes 12 temas diferentes y ellos pueden elegir libremente uno para sobre él redactar la composición.[28]
- Todos los niveles de competencia (principiante, intermedio y avanzado) están representados en el corpus. Es especialmente destacable el hecho de que, para establecer el nivel real de los estudiantes, estos completaron –de forma telemática– el College-Level Placement Test de la Universidad de Wisconsin.[29]
- En cuanto al muestreo, se escoge un diseño transversal estándar, basado en diferentes niveles de competencia.

Principio 3: *Contraste*

- Entre el subcorpus de hablantes nativos de español y el de aprendices.
- Entre dos interlenguas (i.e. dos niveles de competencia).

Principio 4: *Criterios estructurales*. División del corpus en un subcorpus de aprendices (compuesto por tres niveles) y un subcorpus de nativos.

Principio 5: *Etiquetado.* Para la anotación del corpus se emplea el etiquetador *UAM Corpus Tool*, que "deja el fichero de texto puro intacto y crea un nuevo fichero XML que contiene las etiquetas" (Lozano Pozo, 2009, p. 201).

---

[26] "El objetivo del proyecto de investigación WOSLAC es alcanzar el millón de palabras en CEDEL2" (Lozano Pozo, 2009, p. 205).

[27] Una descripción detallada de la aplicación de estos principios al diseño de CEDEL2 se puede encontrar en el artículo "CEDEL2: Corpus Escrito del Español L2" (Lozano Pozo, 2009), del cual hemos extraído el esquema que presentamos aquí.

[28] Los corpus de aprendices escritos suelen utilizar como tarea para la recopilación de los textos la redacción. Ello posee la gran ventaja de que el grado de libertad es máximo y, por tanto, "el grado de proximidad a una elocución natural y no guiada" (Santos Gargallo, 1993, p. 107) también lo es. No obstante, hemos de tener en cuenta que "[e]l uso de una composición como fuente de datos para la investigación siempre plantea el problema de que el alumno evitará, consciente o inconscientemente, aquellas estructuras de las que no se sienta muy seguro, y nunca sabremos si las evitó porque no eran necesarias para la transmisión del mensaje o porque quiso evitarse problemas" (Santos Gargallo, 1993, p. 107).

[29] En el corpus WriCLE también podemos encontrar medidas del nivel de competencia de los estudiantes determinadas por un test de diagnóstico estandarizado. Sin embargo, ICLE, SPLLOC y muchos otros corpus no tienen en cuenta este aspecto, y clasifican a los aprendices "por medio de factores externos como el nivel de dominio que deberían tener de acuerdo con la edad y curso en el que se encuentran. [Sin embargo,] [d]ichos factores no garantizan que los participantes de un corpus sean comparables en términos de dominio lingüístico, ni que los resultados de los análisis sean extrapolables a otros contextos de aprendizaje similares" (Sánchez Rufat, 2015, p. 199).

Principio 6: *Muestra*. Se compilan solo textos completos, aunque difieran considerablemente en el número de palabras que los componen.

Principio 7: *Documentación*. Se incluyen datos sobre cada estudiante, relativos a diferentes aspectos.

Principio 8: *Equilibrio*. En el diseño de un corpus son fundamentales la representatividad y el balance. En el caso de CEDEL2, dado que se trata de un corpus escrito, los resultados solo podrán extrapolarse a la interlengua escrita.

Principio 9: *Tema*. "[L]os temas de redacción están sopesados para dar lugar a un lenguaje lo más representativo posible" (Lozano Pozo, 2009, p. 202).

Principio 10: *Homogeneidad*. Con el objetivo de descartar los textos que no se adecuen a los criterios del corpus (*rogue texts*), los investigadores examinan cada composición antes de etiquetar los datos.

Así pues, este corpus no ha sido diseñado con el propósito de elicitar unas estructuras lingüísticas concretas ni unos determinados elementos léxicos en las redacciones de los estudiantes, sino que se pretende que el contenido obtenido sea variado y representativo de la interlengua de los aprendientes de ELE.[30] Ello ofrece la enorme ventaja de "poder contestar a cualquier pregunta de investigación sobre la adquisición del español" (Sánchez Rufat, 2015, p. 200).

CEDEL2 se encuentra disponible de manera gratuita en la página electrónica http://cedel2.learnercorpora.com/.

CAES (*Corpus de Aprendices del Español*)

CAES es el resultado de un proyecto emprendido por el Instituto Cervantes y llevado a cabo por un equipo de investigación de la Universidad de Santiago de Compostela. La versión 1.0. se publicó en acceso abierto en octubre de 2014 y en marzo de 2018 se ha lanzado la versión 1.1.. Este corpus cuenta con casi 575 000 elementos lingüísticos y está compuesto por pruebas escritas recogidas en un período temporal que abarca desde octubre de 2011 hasta septiembre de 2013. Tras un proceso de desambiguación manual implementado en las 3878 tareas (producidas por 1423 aprendices), estas fueron anotadas en cuanto al plano morfosintáctico de manera automática. Ese etiquetado fue revisado posteriormente de forma manual también, con el objetivo de aumentar su fiabilidad (Parodi, 2015).

El Corpus de Aprendices del Español está disponible en la página electrónica http://galvan.usc.es/caes/search. La interfaz de búsqueda permite filtrar según los siguientes parámetros usados en la configuración del corpus:

- Nivel de español (de A1 a C1)
- Lengua materna del aprendiz (seis: árabe/ chino mandarín/ francés/ inglés/ portugués/ ruso)
- País de residencia
- Edad
- Sexo

Las distintas combinaciones posibles atendiendo a estos filtros favorecen la realización de numerosos y variados estudios que permiten la consecución de muy diversos objetivos.

CAELE (*Corpus de Aprendices de Español como Lengua Extranjera*)

[30] Por tanto, en palabras de Sánchez Rufat (2015, p. 200), "CEDEL2 se considera una base de datos naturales".

Este corpus se recopiló en Chile durante los años 2014 y 2015. La creación y anotación del mismo forma parte de un proyecto cuya investigadora responsable es la Dra. Ferreira Cabrera.

CAELE está formado por 418 textos escritos por 62 estudiantes de los cursos de español general como lengua extranjera de una universidad chilena. Las lenguas maternas de estos aprendientes son: alemán, francés, inglés, portugués, sueco, checo, italiano y ruso. Con relación al nivel de competencia, los sujetos se distribuyeron en dos grupos (A2+ y B1) según los resultados que obtuvieron en el examen de Certificación del Español como Lengua Extranjera (CELE) (Ferreira Cabrera y Elejalde Gómez, 2017).

El equipo de investigación que compiló este corpus lo ha usado en un estudio cuyo objetivo es "determinar los errores más frecuentes y recurrentes en el CAELE" (Ferreira Cabrera y Elejalde Gómez, 2017, p. 516). Con este propósito, han procedido a la anotación de los errores presentes en los textos que componen el corpus. Para ello, siguiendo un criterio lingüístico, han establecido una taxonomía con tres niveles de categorización (gramática, léxico y ortografía) y los siguientes tipos de errores: adición; falsa selección; omisión; forma errónea; elección errónea (Ferreira Cabrera y Elejalde Gómez, 2017).

Para anotar el corpus, Ferreira Cabrera y su equipo han utilizado el programa informático *UAM Corpus Tool* (versión 3.2.), desarrollado por Michael O'Donell en la Universidad Autónoma de Madrid. Este software, de distribución gratuita con fines de investigación, proporciona herramientas que permiten el etiquetado manual, semiautomático y automático. Es destacable el hecho de que ofrece la opción de "etiquetado múltiple", la cual hace posible marcar un mismo error como perteneciente a más de un nivel. Utilizando, como hemos señalado anteriormente, la taxonomía de errores que ellos mismos han elaborado, los creadores de CAELE han diseñado, de forma manual, un esquema de anotación en UAM Corpus Tool, y además han asignado diversos atributos a los textos para poder clasificarlos según las variables del estudio –lengua materna, nivel, etc.– (Ferreira Cabrera y Elejalde Gómez, 2017).

SPLLOC (*Spanish Learner Language Oral Corpus*)

A diferencia de todos los corpus de aprendices de ELE abordados con anterioridad, que estaban compuestos por materiales escritos, SPLLOC es una recopilación de textos orales que se encuentran disponibles para su descarga (tanto los archivos de sonido como su transcripción) en la página http://www.splloc.soton.ac.uk/.

El diseño de SPLLOC incluye varios géneros (narrativo, entrevistas, descripción de fotos / dibujos y debates entre iguales). Asimismo, recoge una importante muestra de habla de cada participante individual (40 – 60 minutos), con distintos interlocutores y también otras tareas. Los datos proceden tanto de estudiantes cuya lengua materna es el inglés como de hablantes nativos de español. Los aprendices de ELE se clasifican en tres niveles: iniciales, intermedios y avanzados (Mendikoetxea, 2014).

Hasta el momento se han emprendido dos proyectos independientes que han utilizado el *Spanish Learner Language Oral Corpus* para estudiar las manifestaciones orales de la interlengua de estudiantes de ELE. El primero de ellos, SPLLOC1, investiga la adquisición de rasgos morfosintácticos fundamentales, como el orden de palabras y los pronombres clíticos. El segundo, SPLLOC2, ahonda en el desarrollo del sistema "tiempo verbal – aspecto" en español como L2 (Mendikoetxea, 2014).

Análisis de errores asistido por ordenador (Computer-aided Error Analysis –CEA–)

El Análisis de Errores asistido por ordenador "es un enfoque de investigación basado en corporas [sic] electrónicos de aprendientes y en los procedimientos de la LC [Lingüística de Corpus] para la identificación, clasificación y descripción de los errores" (Ferreira Cabrera, Elejalde Gómez y Vine Jara, 2014, p. 390). Se trata, pues, de solventar las deficiencias metodológicas del Análisis de Errores (*véase* el apartado *Análisis de Errores (AE)* del presente trabajo) mediante la utilización de herramientas informáticas que aportan rigor y sistematicidad a los estudios. EL CEA se beneficia, por tanto, de los métodos procedentes de la Lingüística de Corpus, "que ayudan a aportar descripciones mejoradas de la lengua del aprendiente" (Sánchez Rufat, 2015, p. 193).

En un artículo en el que, ya en 1998, se introduce la técnica del Análisis de Errores asistido por ordenador, Dagneaux, Denness y Granger explican que, hasta aquel momento, los programas informáticos que ofrecían análisis lingüísticos automáticos estaban diseñados para investigar variedades nativas de la lengua exclusivamente. Por ello, la aplicación de estas herramientas a la investigación sobre textos escritos por aprendices del idioma entrañaba muchas dificultades (sobre todo en el área de la gramática y el estilo), ya que los errores cometidos por los estudiantes difieren considerablemente de los producidos por nativos. Por tanto, se hacía necesario elaborar un sistema que hiciera posible la creación de un instrumento especializado en el análisis de errores de aprendices.

El grupo de lingüistas de corpus de Lovaina desarrolló un sistema de CEA con el objetivo referido al final del párrafo anterior. Se optó por una descripción de los errores basada en criterios puramente lingüísticos, dado que las categorizaciones relativas a las causas o fuentes del error se caracterizan por un alto grado de subjetividad.[31] De este modo fue posible la creación de un sistema de etiquetado de errores jerárquico, en el que las etiquetas constaban de un código principal y una serie de subcódigos. Los siete códigos principales (*mayor category codes*) que se fijaron fueron los siguientes: formal, gramatical, léxico-gramatical, léxico, registro, palabra redundante / omisión de palabra / orden de palabras y estilo (Dagneaux, Denness y Granger, 1998).

Análisis Contrastivo de Interlengua –ACI– (Contrastive Interlanguage Analysis –CIA–)

Mendikoetxea (2014) define el Análisis Contrastivo de Interlengua como un paradigma de investigación que establece comparaciones entre:

a) dos (o más) variedades de interlengua (por ejemplo, L1 español – L2 inglés vs. L1 italiano – L2 inglés)
b) la gramática de la L1 y la de la L2, mediante la comparación de un corpus de nativos con uno de no nativos

Según detalla Sánchez Rufat (2015a), Granger propone el Análisis Contrastivo de Interlengua en 1996, y le otorga una gran relevancia en el marco del proyecto ICLE (*véase* el apartado *Los corpus de aprendices* del presente trabajo). Es reseñable el hecho de que,

> [a] diferencia del estudio contrastivo, que implica normalmente la comparación de dos lenguas, el ACI involucra diferentes variedades de la misma lengua: "involves quantitative and qualitative comparisons between native language and learner language (L1 *vs*. [*versus*] L2) and between different varieties of interlanguage (L2 vs. L2)" (Granger, 2009: 18). (Sánchez Rufat, 2015a, p. 196)

[31] Únicamente un tipo de error de los incluidos en el sistema constituye una excepción en la taxonomía lingüística elegida: los "false friends". El motivo de su adición es el especial interés que el equipo de Lovaina tiene en esta clase de error (Dagneaux, Denness y Granger, 1998).

Finalmente, hemos de destacar que, pese a las fuertes críticas que ha recibido,[32] el Análisis Contrastivo de Interlengua ha experimentado un gran desarrollo dentro de los estudios de interlengua basados en corpus. No obstante, en el caso de la interlengua del español el camino recorrido es todavía corto, como apunta Sánchez Rufat (2015b).

La investigación basada en corpus de aprendices[33] y la Adquisición de Segundas Lenguas (ASL)

Hasta hace relativamente poco tiempo, el área de la ASL basaba sus estudios únicamente en datos experimentales o procedentes de la introspección, lo cual implicaba que los investigadores contaban con un número reducido de participantes y, por tanto, la generalización de los resultados obtenidos no estaba exenta de controversia (Sánchez Rufat, 2015b).

Los expertos en Adquisición de Segundas Lenguas tendían a rechazar los datos de usos lingüísticos naturales –presentes en los corpus– por varias razones, entre las que destacan las dificultades para controlar las variables que pueden afectar a la producción de los estudiantes o la escasa formación de los lingüistas aplicados en el manejo de programas informáticos complejos.[34] Sin embargo, el hecho de contar con una base empírica tan estrecha, llegando en ocasiones a estudios de caso, no permite analizar la lengua colectiva del aprendiz. De acuerdo con Sánchez Rufat (2015b, p. 72), "[e]sto resulta llamativo cuando precisamente las dificultades típicas de un grupo concreto son las que deben (…) interesar en las investigaciones de la interlengua".

Por ello, parece claro que, pese a los inconvenientes que pueda conllevar, el uso de corpus de aprendientes se muestra útil en los estudios de ASL porque constituye una fuente de datos naturales a gran escala. Como leemos en Sánchez Rufat (2015b), diversos autores –entre los que se encuentran Granger (2002, 2012), Myles (2005) y Lozano Pozo y Mendikoetxea (2013)– han defendido esta propuesta y han demostrado las ventajas de implementarla.

Así pues, la investigación basada en corpus de aprendices ha creado un vínculo entre dos disciplinas antes discrepantes: la Lingüística de Corpus y la ASL –Granger (2002); Lozano y Mendikoetxea (2013); Sánchez Rufat (2015b)–. No obstante, pese a la conveniencia de utilizar datos naturales a gran escala para llevar a cabo estudios en el ámbito de la Adquisición de Segundas Lenguas, no hay que perder de vista que

> la ASL siempre requerirá otras fuentes de datos, como los experimentales, los metalingüísticos y los de la introspección (como los que se usan tradicionalmente en la SLA [Second Language Acquisition]), para contrastar los resultados obtenidos del análisis de corpus y para triangular los resultados y obtener así resultados más convincentes (Guilquin y Gries, 2009; Mendikoetxea y Lozano, 2012, 2013; citados en Sánchez Rufat, 2015b, p. 78).

Por último, hemos de señalar que los hallazgos encontrados en estudios basados en corpus de aprendices han tenido fundamentalmente una aplicación didáctica (*véase* el apartado *La investigación basada en corpus de aprendices y la Enseñanza de Lenguas Extranjeras* del presente trabajo),[35] mientras que poco se ha avanzado en la utilización

[32] Para profundizar en las razones motivadoras de dichas críticas, *véase* Granger (2009).
[33] Otro término empleado para designar este campo de estudio es "investigación de Corpus de Aprendientes Informatizados" (Sánchez Rufat, 2015, p. 73).
[34] Con relación a este aspecto, Lozano (2015) enfatiza la necesidad de que los especialistas en investigación en corpus de aprendices se esfuercen por aplicar análisis de estadística inferencial a sus resultados para dotar a sus afirmaciones y conclusiones de un respaldo sólido.
[35] No obstante, Johansson (2009) recuerda la necesidad de contar con más estudios sistemáticos sobre la efectividad de los corpus en la enseñanza de lenguas. Según indica este autor, pese a la existencia de

de las teorías y los resultados empíricos obtenidos en el ámbito de la Adquisición de Segundas Lenguas como base para llevar a cabo una correcta interpretación de dichos hallazgos (Lozano, 2015). Por ello, sería interesante volver la vista hacia el bagaje investigador existente en ASL para superar la naturaleza descriptiva y exploratoria que ha caracterizado hasta ahora los trabajos basados en corpus de aprendices y que estos comiencen a profundizar en la explicación y la interpretación de los resultados que presentan (Myles, 2005; Granger, 2009; Díaz Negrillo y Thompson, 2013; Lozano y Mendikoetxea, 2013; Lozano, 2015).

La investigación basada en corpus de aprendices y la Enseñanza de Lenguas Extranjeras

Aunque la creación de corpus de aprendices es una práctica bastante reciente, la investigación en Enseñanza de Lenguas Extranjeras (en concreto, en Inglés como Lengua Extranjera –ILE–) lleva muchos años haciendo uso de corpus de nativos de la lengua meta para diseñar materiales y mejorar la metodología docente (Granger, 2002).[36]

Dentro del ámbito del diseño de materiales, el campo en el que han tenido lugar los avances más significativos gracias a la utilización de corpus de nativos es el de la producción de diccionarios de ILE. La información contenida en estos se ha enriquecido notablemente, incorporando datos relativos a patrones gramaticales, colocaciones (i.e. afinidades léxicas), estilo, frecuencias, etc. Asimismo, otros materiales como las gramáticas o los libros de texto de ILE también se han beneficiado de las posibilidades que ofrecen los corpus de nativos. El área de la metodología docente, por su parte, ha ampliado el abanico de recursos didácticos disponibles mediante la adición de ejercicios basados en concordancias, obtenidas estas al realizar búsquedas en corpus de hablantes de inglés como lengua materna (Granger, 2002).

Sin embargo, como indica esta autora, los corpus de nativos no proporcionan ninguna indicación sobre las dificultades de aprendizaje que experimentan los estudiantes en general o determinados grupos de estudiantes en particular. Así pues, resulta de especial relevancia que el uso de estos se complemente con el análisis de los datos que proporcionan los corpus de aprendices. Además, sería muy interesante también contar con corpus bilingües que contuvieran textos escritos en la lengua meta y en el idioma materno de los estudiantes, ya que la utilización conjunta de todos estos recursos arrojaría luz sobre el proceso de aprendizaje.

Esta perspectiva propuesta por Granger (2002) repercutiría positivamente tanto en la enseñanza del vocabulario como de la gramática. De igual modo, el diseño de diversos materiales (como los diccionarios monolingües, los bilingües y los programas CALL –Computer Assisted Language Learning–) se vería favorecido. En cuanto a la metodología docente, pese al riesgo que supone exponer al alumnado a datos erróneos, Granger (2002) defiende el uso de los corpus de aprendices en el aula en los dos contextos siguientes: “form-focused instruction” (p. 26) –con el objetivo de hacer conscientes a los estudiantes de las diferencias formales existentes entre su lengua materna y la lengua meta– y “learning-driven data” (p. 26) –enfoque en el cual el alumnado trabaja con y sobre su propia producción–.

Por último, destacaremos la clasificación establecida por Granger (2009, p. 20), quien divide los corpus de aprendices en dos categorías: los de uso pedagógico aplazado –“delayed pedagogical use (DPU)”– y los de uso pedagógico inmediato –(“immediate pedagogical use (IPU)”–. Los primeros se recopilan con un propósito investigador,

numerosas publicaciones que ilustran cómo puede utilizarse un corpus en el aula, no se ha ahondado en la evaluación de los beneficios de este enfoque.

[36] Para desarrollar la presente sección nos basaremos principalmente en el capítulo de libro “A Bird’s eye view of learner corpus research” publicado por Sylviane Granger en 2002.

mientras que los segundos son compilados con vistas a su uso en clase para trabajar con los mismos estudiantes que han producido los textos que los componen –*véanse* también Johansson (2009) y Díaz Negrillo y Thompson (2013)–.

Los manuales de ELE dirigidos a estudiantes de L1 inglés

Encontramos en el mercado manuales de enseñanza de ELE dirigidos a estudiantes de una determinada L1, los cuales abordan aspectos del español que pueden resultar problemáticos a los aprendices que poseen ese idioma materno concreto.

Con el propósito de presentar una breve descripción de una muestra de estos manuales, hemos seleccionado varios que tienen como destinatarios a estudiantes nativos de inglés:

- Bueso Fernández, Isabel y Pilar Casamián Sorrosa (2010 [2001]) *Diferencias de usos gramaticales entre el español y el inglés*, Madrid: Edinumen, Colección Temas de español, gramática contrastiva
  Esta obra se concibe como material complementario y está dirigida a estudiantes de niveles inicial e intermedio. Aborda "los aspectos morfológicos, sintácticos y funcionales más importantes de ambas lenguas, tanto los que se asemejan como los que difieren" (Bueso Fernández y Casamián Sorrosa, 2010, p. 7).
- Capón, María Luisa y Manuela Gil (2003) *Dificultades del español para hablantes de inglés*, Madrid: SM, Colección Prácticos ELE
  En este libro encontramos una exposición y descripción de los errores más frecuentes producidos por los estudiantes de ELE que poseen como L1 el inglés. Incluye asimismo claves y fórmulas para corregir los errores y ejercicios de consolidación.
- Fernández Agüero, María (2007) *Español para hablantes de inglés*, Madrid: SGEL ele, Colección Contrastes
  Se trata de un libro de autoaprendizaje que puede utilizarse también como material complementario. En él aparecen indicaciones útiles que advierten a los estudiantes angloparlantes de nivel B1 sobre los errores más comunes que suelen cometer los nativos de inglés cuando aprenden español como lengua extranjera. Ofrece ejemplos en el idioma meta y en la L1 del estudiante, así como actividades para practicar los contenidos.
- Muñoz-Basols, Javier, Marianne David y Olga Núñez Piñeiro (2010) *Speed up your Spanish. Strategies to avoid common errors*, London / New York: Routledge
  Esta obra, de mayor extensión que las tres anteriores, está diseñada para estudiantes de nivel intermedio. Constituye un excelente y completo manual que abarca una gran diversidad de contenidos, desde los falsos amigos hasta las expresiones idiomáticas, poniendo el foco en las diferentes categorías gramaticales en las que los nativos de lengua inglesa encuentran escollos a la hora de aprender español. Además, posee una página electrónica complementaria en la que se pueden encontrar ejercicios adicionales y archivos de audio.

A modo de conclusión

En este trabajo hemos presentado una revisión bibliográfica centrada en las publicaciones que versan sobre los corpus de aprendices en general y los de Español como Lengua Extranjera en particular. Dado que el estudio de los errores presentes en las producciones de los estudiantes constituye una de las principales áreas abordadas por la investigación en corpus de aprendices, hemos comenzado realizando un recorrido por la historia de las investigaciones en Lingüística Contrastiva desde sus orígenes, para después poner de manifiesto que la ayuda proporcionada en la actualidad por los ordenadores ha solventado problemas que se detectaron en los estudios de la segunda

mitad del siglo XX (especialmente en los que implementaban el modelo del Análisis de Errores).

A continuación, nos hemos centrado en la investigación basada en corpus de aprendices. Hemos descrito cinco corpus constituidos por textos de estudiantes de Español como Lengua Extranjera: CORANE, CEDEL2, CAES, CAELE y SPLLOC. Más adelante hemos abordado dos tipos de análisis que hacen uso de corpus electrónicos: el Análisis de errores asistido por ordenador y el Análisis Contrastivo de Interlengua. Después hemos indagado en las publicaciones que han profundizado en la relación entre la investigación basada en corpus de aprendices y la Adquisición de Segundas Lenguas, por un lado, y entre aquella y la Enseñanza de Lenguas Extranjeras, por otro.

Por último, hemos dedicado una sección a la descripción de varios manuales de ELE dirigidos a estudiantes que poseen el inglés como lengua materna.

* * *

La investigación basada en corpus de aprendices es una rama de estudio iniciada hace aproximadamente 30 años (Sylviane Granger fundó el Centro de Lingüística de Corpus del Inglés de la Universidad Católica de Lovaina en 1990 [37] – https://uclouvain.be/en/research-institutes/ilc/cecl–). Los avances tecnológicos han posibilitado que, pese a su corta trayectoria, haya experimentado un gran desarrollo desde principios del siglo XXI,[38] y las perspectivas de crecimiento son favorables. No en vano, las ventajas que proporcionan los corpus de aprendientes al estudio de la interlengua son evidentes:

> Los corpus informatizados permiten aumentar el tamaño del material analizado –una investigación a gran escala puede revelar rasgos de uso que hayan escapado a lingüistas que emplearan la intuición o una pequeña cantidad de muestras– y la variedad, normalmente producto de una gran cantidad de participantes (Sánchez Rufat, 2015b, p. 77).

No obstante, pese a los numerosos aspectos positivos que el uso de corpus de aprendices ofrece, hemos de adoptar un punto de vista realista con relación a las posibilidades que proporcionan (Granger, 2009). Como indica Johansson (2009), los corpus no pueden reemplazar a la comunicación natural ni sustituir al profesorado, pero si se usan de forma racional constituyen una herramienta de aprendizaje eficaz.

Referencias

Alexopoulou, A. (2006). Los criterios descriptivo y etiológico en la clasificación de los errores del hablante no nativo: una nueva perspectiva. *Porta Linguarum: Revista Internacional de Didáctica de Las Lenguas Extranjeras*, *5*, pp. 17-36.

Ballier, N., Díaz-Negrillo, A. y Thompson, P. (2013). Introduction. En *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 3-7). Amsterdam / Philadelphia: John Benjamins Publishing Company.

[37] Como señala Sánchez Rufat, otra fecha relevante es 1998, año en el que Granger edita *Learner English on Computer*, "una colección de artículos pioneros sobre la lengua del aprendiente, basada en el ICLE" (Sánchez Rufat, 2015, p. 195).

[38] Este gran desarrollo se ve reflejado en las numerosas publicaciones sobre la interlengua de aprendices de inglés que han aparecido en los últimos años. Sin embargo, en el caso del español, la investigación basada en corpus de aprendices tiene aún un largo camino por recorrer –Lozano (2015), Sánchez Rufat (2015)–.

Baralo Ottonello, M. (2009). A propósito del Análisis de errores: una encrucijada de teoría lingüística, teoría de adquisición y didáctica de lenguas. *Revista Nebrija de Lingüística Aplicada a La Enseñanza de Las Lenguas*, *5*(3), pp. 27-31.

Belda Torrijos, M. (2015). *Errores del uso del español en estudiantes extranjeros: creación de una taxonomía y su etiquetado* (Tesis Doctoral). Valencia: Universitat Politècnica de València.

Bueso Fernández, I. y Casamián Sorrosa, P. (2010). *Diferencias de usos gramaticales entre el español y el inglés. Gramática contrastiva*. Madrid: Edinumen.

Capón, M. L. y Gil, M. (2003). *Dificultades de español para hablantes de inglés*. Madrid: SM.

Centro Virtual Cervantes. (n.d.). *Diccionario de términos clave de ELE*. Recuperado de http://cvc. cervantes.es/ensenanza/biblioteca_ele/diccio_ele/indice.htm

Cestero Mancera, A. M., Penadés Martínez, I., Blanco Canales, A., Camargo Fernández, L. y Simón Granda, J. F. (2002). Corpus para el análisis de errores de aprendices de E/LE (CORANE). En *Actas del XII Congreso Internacional de ASELE: tecnologías de la información y de las comunicaciones en la enseñanza de la E/LE, Valencia, 2001.* (pp. 527-534). Valencia: Universidad Politécnica de Valencia.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Dagneaux, E., Denness, S. y Granger, S. (1998). Computer-aided error analysis. *System*, *26*(2), pp. 163-174.

de Alba Quiñones, V. (2009). El análisis de errores en el campo del español como lengua extranjera. Algunas cuestiones metodológicas. *Revista Nebrija de Lingüística Aplicada a La Enseñanza de Las Lenguas*, (5), pp. 1-16.

Fernández Agüero, M. (2007). *Español para hablantes de inglés*. Madrid: SGEL.

Fernández López, S. (1995). Errores e interlengua en el aprendizaje del español como lengua extranjera. *Didáctica*, *7*, pp. 203-216.

Ferreira Cabrera, A. y Elejalde Gómez, J. (2017). Análisis de errores recurrentes en el Corpus de Aprendices de Español como Lengua Extranjera, CAELE. *Revista Brasileira de Linguística Aplicada*, *17*(3), pp. 509-538.

Ferreira Cabrera, A., Elejalde Gómez, J. y Vine Jara, A. (2014). Análisis de Errores Asistido por Computador basado en un Corpus de Aprendientes de Español como Lengua Extranjera. *Revista Signos: Estudios de Lingüística*, *47*(86), pp. 385-411.

Granger, S. (2002). A Bird's-eye view of learner corpus research. En *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3-33). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Granger, S. (2012). How to Use Foreign and Second Language Learner Corpora. En *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 7-29). Chichester: Wiley-Blackwell.

Granger, S., Gilquin, G. y Meunier, F. (2015). Introduction: learner corpus research – past, present and future. En *The Cambridge Handbook of Learner Corpus Research* (pp. 1-6). Cambridge: Cambridge University Press.

Johansson, S. (2009). Some thoughts on corpora and second-language acquisition. En *Corpora and Language Teaching* (pp. 34-44). Amsterdam / Philadelphia: John Benjamins Publishing Company.

López Salinas, M.P. (2001). Estudio y análisis de errores de la interlengua de español para anglófonos. En *Interferencias, cruces y errores* (pp. 101-117). Madrid: SGEL.

Lozano Pozo, C. (2009). CEDEL2: Corpus Escrito del Español L2. En *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente* (pp. 197-212). Almería: Universidad de Almería.

Lozano Pozo, C. (2015). Learner corpora as a research tool for the investigation of lexical competence in L2 Spanish. *Journal of Spanish Language Teaching*, *2*(2), pp. 180-193.

Lozano Pozo, C. y Mendikoetxea, A. (2013). Learner corpora and second language acquisition: the design and collection of CEDEL2. En *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 65-100). Amsterdam / Philadelphia: John Benjamins Publishing Company.

McEnery, T., Xiao, R. y Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London & New York: Routledge.

Mendikoetxea, A. (2014). Corpus-based Research in Second Language Spanish. En *The Handbook of Spanish Second Language Acquisition* (pp. 11-29). Oxford: Wiley-Blackwell.

Muñoz-Basols, J., David, M. y Núñez Piñeiro, O. (2010). *Speed up your Spanish. Strategies to avoid common errors*. London / New York: Routledge.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, *21*(4), pp. 373-391.

Parodi, G. (2015). Corpus de aprendices de español (CAES). *Journal of Spanish Language Teaching*, *2*(2), pp. 194-200.

Ruiz Martín, A. (2006). *Los efectos de la instrucción enfocada al procesamiento del input en la adquisición de la concordancia de género de los sustantivos 'problemáticos', con sus modificadores adjetivos, por estudiantes angloparlantes de ELE* (Proyecto de investigación tutelada). Universidad Nebrija, Biblioteca 2006. Número 5. Primer semestre 2006.

Sánchez Rufat, A. (2015a). Análisis contrastivo de interlengua y corpus de aprendientes: precisiones metodológicas. *Pragmalingüística*, *23*, pp. 191-210.

Sánchez Rufat, A. (2015b). La investigación de corpus de aprendientes y el desarrollo de los estudios de la interlengua del español. *Language Design*, *17*, pp. 57-84.

Sánchez Rufat, A. y Jiménez Calderón, F. (2013). Apreciaciones sobre la cuestión de la norma en el análisis de la interlengua. *Normas: Revista de Estudios Lingüísticos Hispánicos*, *3*, pp. 183-204.

Santos Gargallo, I. (1993). *Análisis contrastivo, análisis de errores e interlengua en el marco de la Lingúística Contrastiva*. Madrid: Editorial Síntesis.

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

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*(3), pp. 209-231.

Torijano, J. A. (2006). Lo que nos enseñan los errores. *Signum: Estudos Da Linguagem*, *9*(1), pp. 141-205.

Varón López, A. (2008). Fosilización y adquisición de segundas lenguas (ASL). *Investigación en Humanidades / Kanagawa University Humanities Association Ed.*, *166*, pp. 101-129.

Vez Jeremías, J. M. (2004). Aportaciones de la lingüística contrastiva. En *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2)/ lengua extranjera (LE)* (pp. 147-163). Madrid: SGEL.

Zimny, A. (2016). *Análisis de errores en la adquisición del artículo español por alumnos polacos de ELE* (Tesis Doctoral). Madrid: Universidad Complutense de Madrid.