LA INVESTIGACIÓN DE CORPUS DE APRENDIENTES BASADA EN EL ANÁLISIS CONTRASTIVO DE LA INTERLENGUA: EL CASO DE *DAR*

Anna Rufat Sánchez

UJa EDITORIAL

Rufat Sánchez, Anna
La investigación de corpus de aprendientes basada en el análisis contrastivo de la interlengua : el caso de DAR / Anna Rufat Sánchez. -- Jaén : Editorial Universidad de Jaén, 2019. -- (Lingüística. Enfoque : Lingüística aplicada ; 1)

192 p. ; 17 x 24 cm
ISBN 978-84-9159-221-1
1. Corpus (Lingüística) 2. Análisis Lingüístico I. Jaén. Editorial Universidad de Jaén, ed. II.Título
81-13

Esta obra ha superado la fase previa de evaluación externa realizada por pares mediante el sistema de doble ciego

Colección: Lingüística
Director: Ventura Salazar García

Serie: *Enfoque: Lingüística aplicada, 1*
Coordinador de la serie: Antonio Bueno González

© Anna Rufat Sánchez
© Universidad de Jaén
Primera edición, diciembre 2019
ISBN: 978-84-9159-221-1
Depósito Legal: J-977-2019

Edita
Editorial Universidad de Jaén
Vicerrectorado de Proyección de la Cultura y Deporte
Campus Las Lagunillas, Edificio Biblioteca
23071 Jaén (España)
Teléfono 953 212 355
web: editorial.ujaen.es

editorial@ujaen.es

Diseño
José Miguel Blanco. www.blancowhite.net

Imprime
Gráficas «La Paz» de Torredonjimeno, S. L.

Impreso en España/*Printed in Spain*

«Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos, www.cedro.org) si necesita fotocopiar, escanear o hacer copias digitales de algún fragmento de esta obra».

Índice

Pág.

Introducción XIII

1. La investigación sobre la lengua del aprendiente basada en corpus 1
1.1. *Estudios de la lengua del aprendiente anteriores al uso de corpus* 2
1.1.1. El análisis contrastivo 2
1.1.2. El análisis de errores 3
1.1.3. La ASL y el problema de los datos 7
1.2. *La investigación de corpus de aprendientes* 8
1.2.1. El desarrollo de los corpus de aprendientes 11
1.3. *Estudios sobre la lengua del aprendiente del español: propósitos y alcance de este estudio* 17

2. La competencia léxica y el conocimiento del verbo 23
2.1. *Los rasgos morfológicos* 24
2.2. *La naturaleza semántica* 25
2.3. *El significado y sus relaciones paradigmáticas y sintagmáticas* 28
2.4. *La interfaz léxico-sintaxis* 30
2.5. *Los rasgos sociolingüísticos* 34
2.6. *Los rasgos de la competencia léxica del verbo* 36

3. La naturaleza semántica y la proyección sintáctica del verbo *dar* 39
3.1. *Consideraciones sobre la naturaleza semántica de* dar 40
3.1.1. Hacia una definición de *dar* que se visualice en la sintaxis 40
3.1.1.1. Los contenidos aspectuales y las idiosincrasias léxicas en el significado de *dar*: la trayectoria y la meta en la transferencia y en la creación .. 41
3.1.1.2. La concordancia léxica y la redundancia 45
3.1.2. Descripción de los sentidos de *dar* 49
3.2. *Usos y estructuras del verbo* dar 54

4. Los procedimientos para el análisis 61
4.1. *El aspecto de la norma* 61
4.2. *La obtención de los datos* 64
4.2.1. El CEDEL2 64
4.2.2. Antconc 68
4.3. *Procedimientos para el análisis de los datos* 72
4.3.1. Cuestiones preliminares 72

4.3.2. Estructuras consideradas en el análisis 73
4.3.3. Fuentes para contrastar los datos 74
4.3.3.1. Introspección 75
4.3.3.2. Otros corpus textuales 76
4.3.3.3. Diccionarios combinatorios 77
4.3.3.4. Obras de consulta generales de carácter panhispánico 78
4.3.4. Criterios para determinar el grado de aceptabilidad de los datos 78
4.3.5. Criterios para determinar la inaceptabilidad de los datos 83
4.3.5.1. Localización del error 84
4.3.5.1.1. Error relevante frente a error no relevante 86
4.3.5.2. Descripción del error 88
4.3.5.2.1. El modelo de funcionamiento del lexicón utilizado en la descripción del error: el lexicón y el sistema computacional. 88
4.3.5.2.2. Error léxico y error computacional 90
4.3.5.3. Explicación del error 94

5. El uso del verbo *dar* en el corpus 97
5.1. *Relaciones del índice de frecuencia entre los dos corpus. Análisis cuantitativo y cualitativo* 98
5.1.1. Frecuencia total del lema *dar*: ¿tienden los HNN a utilizar este lema más o menos que los HN? 100
5.1.1.1. Análisis cuantitativo 100
5.1.1.2. Análisis cualitativo 101
5.1.2. Construcciones sintácticas del verbo y su frecuencia 102
5.1.2.1. Análisis cuantitativo 103
5.1.2.2. Análisis cualitativo 105
5.1.2.3. Interpretación de las relaciones de frecuencia 106
5.1.2.3.1. Alta frecuencia en HNN vs. alta frecuencia en HN 107
5.1.2.3.2. Baja frecuencia en los HNN vs. baja frecuencia en los HN 112
5.1.2.3.3. No frecuencia en HNN vs. baja frecuencia en HN 119
5.1.2.3.4. Síntesis del análisis de las relaciones de frecuencias 120
5.1.3. Sustantivos y clases léxicas con las que se combina *dar* 121
5.1.3.1. El uso de *dar* como verbo de transferencia 123
5.1.3.1.1. Análisis cuantitativo 123
5.1.3.1.2. Análisis cualitativo 125
5.1.3.2. El uso de *dar* como verbo de creación 126
5.1.3.2.1. Análisis cuantitativo 126
5.1.3.2.2. Análisis cualitativo 129
5.2. *Aceptabilidad e inaceptabilidad de los datos no nativos* 130
5.2.1. Frecuencia de los datos aceptable-naturales, aceptable-restringidos einaceptables 131
5.2.2. Localización del error 132
5.2.3. Descripción y explicación del error en el núcleo verbal 132
5.2.4. El error y el acierto en las estructuras analizadas 138
5.2.5. Descripción y explicación del error en el complemento 140

5.2.6. Descripción y explicación del error en el sujeto 143
5.2.7. El error en el sintagma..... 143
5.3. *Observaciones sobre el uso de dar* 145
5.3.1. Diferencias y semejanzas por estructuras..... 145
5.3.2. Diferencias y semejanzas de los usos de dar (V + N) en sus dos significados 146
5.3.3. El papel de la transferencia de la L1 147

6. Conclusiones 149
6.1. *Sobre la dificultad de uso de* dar..... 149
6.2. *Implicaciones del análisis y sus resultados* 151
6.2.1. Aplicaciones didácticas 153
6.2.1.1. Materiales didácticos 154
6.2.1.2. Contenidos léxicos..... 156
6.2.1.3. Procedimientos de enseñanza 157
6.3. *Los estudios de la interlengua del español en el futuro* 162

Referencias bibliográficas 165

Lista de siglas

AC	Análisis contrastivo
ACI	Análisis contrastivo de la interlengua
AE	Análisis de errores
ASL	Adquisición de segundas lenguas
CAI	Corpus de aprendientes informatizado
CLCI	Centro de Lingüística de Corpus del Inglés
CEDEL2	Corpus Escrito de Español como Segunda Lengua
CD	Complemento directo
CI	Complemento indirecto
CVA	Construcción con verbo de apoyo
CVP	Construcción con verbo que actualiza su sentido pleno
DPD	Diccionario panhispánico de dudas
ELE	Español como lengua extranjera
HN	Hablante nativo
HNN	Hablante no nativo
ICLE	International Corpus of Learner English
L1	Lengua materna
L2	Lengua meta
LE	Lengua meta
LC	Lingüística de corpus
LINDSEI	Louvain International Database of Spoken English Interlanguaje
MCER	Marco común europeo de referencia para las lenguas
N	Nombre
NGLE	Nueva gramática de la lengua española
SN	Sintagma nominal
SPLLOC	Spanish Learner Language Oral Corpora
V	Verbo
SP	Sintagma preposicional

Introducción

Este libro constituye un análisis del uso del verbo *dar* en la interlengua del español de un grupo representativo de aprendientes anglófonos a través de una investigación de corpus de aprendientes basada en el método del análisis contrastivo de la interlengua (Granger, 1996a, 2015). Combina, por lo tanto, perspectivas y planteamientos de la adquisición de segundas lenguas, la lingüística aplicada y la lengua española.

El propósito del libro es, por un lado, familiarizar a investigadores y profesores de español como segunda lengua con los principios fundamentales de la investigación de corpus informatizados de aprendientes a través de una metodología diseñada expresamente para el análisis de la interlengua que, al mismo tiempo, es consistente, informativa, flexible y reusable; y, por otro lado, mostrar que nuestro conocimiento acerca de cómo se adquiere el español como L2 puede verse ampliado gracias a este enfoque metodológico.

Para este análisis, se ha seleccionado un elemento léxico porque son todavía muy pocos los estudios de interlengua del español que tienen como objeto de estudio este componente, pese a la revalorización que ha experimentado en los últimos cuarenta años en las teorías lingüísticas, psicolingüísticas y pedagógicas. Dentro del léxico, el verbo es un área muy importante de la estructura de cualquier lengua, ya que contiene concreciones argumentales que provocan que el significado de la oración se asimile al del verbo que la proyecta. Esto lo convierte en un área fundamental de la estructura de cualquier lengua, susceptible de generar errores (Housen, 2002: 78).

En concreto, el verbo *dar* es uno de los más representativos del grupo de verbos frecuentes del español y, al mismo tiempo, forma parte del grupo de palabras "pequeñas" (del inglés *smallwords*), que parecen desempeñar un papel central en el éxito comunicativo (Sinclair, 1991; Hasselgren, 2002), lo que lo sitúa como un buen punto de partida para evaluar la lengua de este grupo de hablantes. Parece evidente que el hecho de que un elemento sea frecuente en la lengua conlleva que el aprendiente esté muy expuesto a él y, por lo tanto, resulte más fácil de recordar (Nation, 2001; Nagini, 2008; Schmitt, 2010; Schmitt y Redwood, 2011). No obstante, las características atribuidas a los verbos frecuentes (Viberg, 1996), así como los datos observados en el aula, sumados a los debates actuales sobre la ASL, conducen a pensar que este verbo es un elemento difícil de manejar y que así lo perciben los profesores y los alumnos.

Algunas de las hipótesis de partida que explicarían esta dificultad sentida son las siguientes. En primer lugar, este verbo polisémico participa en construcciones sintácticas

muy variadas; una variación que es tanto intralingüística (piénsese en *dar por*, *dar en*, *dar a*, *dar con*, *dar contra*, *dar que,* seguidos de nombres, adjetivos, infinitivos, conjunciones, entre otras muchas posibilidades) como interlingüística, ya que las estructuras difieren de una lengua a otra; esto es, no hay una correspondencia plena, en el caso que nos ocupa, entre los usos de *dar* y los de *give*. En segundo lugar, la construcción verbo + nombre (V + N), que desde una perspectiva formal es sencilla, puede resultar compleja desde una perspectiva semántica, pues el verbo coaparece con una gran variedad de nombres. En tercer lugar, si nos detenemos en esta estructura V + N, llaman nuestra atención las *construcciones con verbo de apoyo* (como *dar un salto* o *dar un paseo*) por su complejidad. Estas combinaciones son opacas, comparadas con las construcciones en las que *dar* actualiza un significado pleno —aunque figurado—, próximo a "proporcionar" o "hacer llegar" (como *dar un consejo*); las primeras se consideran más opacas porque es difícil rastrear en ellas el vínculo que hay entre el significado del verbo de apoyo —próximo a "hacer" o "ejecutar"— y el significado básico o físico a partir del cual se proyectan todos los figurados (el dar, de *dar un lápiz*).

Estas tres hipótesis acerca de la dificultad del aprendizaje de este verbo están planteadas en torno a la interfaz léxico-sintaxis, que es uno de los centros de interés reciente en la ASL por considerarse causa potencial de desviaciones en el desarrollo de la interlengua. Así pues, partir de unas hipótesis que establecen un vínculo con los debates actuales sobre la ASL y sus implicaciones en el desarrollo de la lengua de los aprendientes constituye un buen motor para esta investigación que arranca de las siguientes preguntas cuyas respuestas nos han permitido contrastar las hipótesis anteriores:

1. ¿Cuáles son las semejanzas y diferencias más destacadas entre el español de los hablantes no nativos y el español de los hablantes nativos con respecto al uso del verbo *dar*? En relación con ello:

1.1. ¿Tienden los no nativos a utilizar este lema más o menos que los nativos?

1.2. ¿Qué estructuras sintácticas utiliza cada grupo, y cuál es la relación del índice de frecuencias?

1.3. ¿Con qué palabras o tipos de palabras combinan los nativos y los no nativos el verbo *dar*?

2. ¿Es un verbo seguro para el no nativo o, por el contrario, es una fuente de problemas fruto de las diferencias interlingüísticas y complejidades intralingüísticas?

3. ¿Qué papel desempeña la L1 en el uso de este verbo? ¿Y qué presencia tiene el factor intralingüístico en la adquisición de este verbo?

En definitiva, como se desprende de las cuestiones anteriores, en este análisis se tienen en cuenta las cuestiones ortográficas, morfológicas, gramaticales, léxicas (asociaciones y colocaciones) y sociolingüísticas del verbo; y en concreto, se presta especial atención a la naturaleza semántica de *dar* y a sus proyecciones sintácticas.

La necesidad de este estudio se ve respaldada por la escasez de trabajos que se centran en este aspecto de la interlengua del español, además de por las discrepancias entre los investigadores sobre verbos frecuentes de la interlengua del inglés: unos hablan de sobreutilización y otros de infrautilización del verbo; unos de verbo seguro y otros de foco de errores; unos subrayan que la influencia de la L1 no es muy importante y, en cambio, otros afirman que el léxico es el nivel más susceptible de ser afectado por la transferencia. Las

conclusiones, además, proporcionarán nuevos caminos para la enseñanza y el aprendizaje del léxico del español basados en las investigaciones de corpus a través de las implicaciones pedagógicas del análisis de interlengua propuesto.

El libro se organiza en cuatro partes principales para cubrir cada uno de los objetivos planteados anteriormente. En la primera sección (capítulo 1) se lleva a cabo una revisión de los estudios sobre la lengua del aprendiente de español que se centra en los corpus de aprendientes informatizados; se revisan los corpus más destacados y los estudios de corpus de español L2 sobre la competencia léxica, y se subrayan los aspectos metodológicos más característicos y las limitaciones de las investigaciones. La segunda sección (capítulos 2 y 3) está dedicada al conocimiento del verbo, en general, y al examen del verbo *dar*, en particular, a través de su naturaleza semántica y su proyección sintáctica. Se propone una descripción de los diferentes significados que actualiza el verbo y de las diferentes estructuras sintácticas en las que participa; por último, con la intención de desarrollar herramientas eficaces para la categorización de los datos analizados del corpus y para facilitar su explicación, se recoge una propuesta formalizada de los aspectos que integran la competencia léxica de este verbo. En la tercera sección (capítulo 4) se describe la metodología de la investigación; tras referir las herramientas empleadas y describir en profundidad los dos corpus comparados y analizados en este trabajo (un corpus nativo y un corpus no nativo de L1 inglés, procedentes del CEDEL2), se presentan los procedimientos seguidos en el análisis de frecuencias y en el de errores; para ello se plantea la cuestión de la selección de la norma en los estudios de interlengua y se delimita el estándar de referencia para el análisis: se propone la aplicación de un concepto de norma que integra un enfoque descriptivo y otro más regulador (prescriptivo). La última sección (capítulo 5), en fin, está dedicada al análisis de la producción de *dar* en el corpus nativo y en el no nativo, que se comparan desde un enfoque cualitativo y cuantitativo. Se atiende, asimismo, al papel de la transferencia de la lengua materna de los aprendientes. En la conclusión (capítulo 6) se discuten los retos y las oportunidades de la futura investigación de corpus de aprendientes de español, así como de la práctica pedagógica basada en los resultados de los análisis de corpus.

1

La investigación sobre la lengua del aprendiente basada en corpus

El principal interés de la investigación en el ámbito de estudio de la adquisición de segundas lenguas (ASL) es encontrar los principios subyacentes que regulan la construcción e interpretación de las estructuras lingüísticas y comunicativas en los diferentes estadios de adquisición de la segunda lengua (L2)[1]. De esta manera, la ASL se plantea como objetivo construir modelos de representación de la lengua de los hablantes no nativos (HNN) en estadios particulares de la adquisición, atendiendo a los factores que limitan o favorecen el desarrollo de dicha adquisición (Larsen-Freeman y Long, 1994 [1991]; Lozano y Mendikoetxea, 2013). La principal fuente de datos para describir el proceso de adquisición y los factores que afectan a este proceso es la propia lengua de los HNN, ya sea producida de manera natural, a través de procedimientos experimentales, o de juicios metalingüísticos (Granger, 2002; Lozano y Mendikoetxea, 2013: 2). Por lo tanto, el éxito de la investigación en ASL depende de la validez y fiabilidad de estos procedimientos de obtención de datos (Lozano y Mendikoetxea, 2013: 1).

Uno de dichos procedimientos se halla en la base del área de investigación lingüística conocida como *investigación de corpus de aprendientes*, que es el resultado de la interrelación de dos disciplinas, hasta hace pocos años desvinculadas: la lingüística de corpus (LC) y la ASL (Granger, 2002: 4). El desarrollo de los corpus de aprendientes informatizados y el de la investigación de la interlengua –o lengua del aprendiente– basada en corpus ha supuesto un gran avance en los estudios de ASL. Mientras que los trabajos anteriores –basados en otros procedimientos de obtención de datos– eran limitados en cuanto al número de sujetos estudiados para poder controlar las variables que afectan a la producción del aprendiente, los nuevos corpus informatizados permiten trabajar con más y mejor calidad de datos de lengua natural. Así pues, poder analizar amplios corpus de aprendientes que estén bien diseñados, de acuerdo con criterios rigurosos (Sinclair, 2005), proporciona una base empírica sólida para la descripción de la interlengua.

[1] Los términos *segunda lengua* y *lengua extranjera* (LE) se emplean aquí indistintamente para aludir a la lengua meta del aprendiente.

A continuación, realizamos un somero repaso por los distintos estadios metodológicos que ha conocido el estudio de la lengua del aprendiente, para desembocar en la investigación de corpus informatizados. Aludimos, después, al papel que estos planteamientos han desempeñado en el ámbito de la interlengua del español, todo lo cual nos sirve como justificación del estudio que llevamos a cabo en el presente trabajo.

1.1. Estudios de la lengua del aprendiente anteriores al uso de corpus

1.1.1. *El análisis contrastivo*

La lengua del aprendiente se convierte en objeto de atención entre los años 40 y 60 del siglo XX, cuando surgen los primeros estudios de adquisición –a medio camino entre los estudios lingüísticos y psicolingüísticos– en medio del debate sobre la teoría del aprendizaje lingüístico y el consiguiente enfrentamiento entre conductistas y cognitivistas (Pastor Cesteros, 2004: 99). Estos estudios orientan por primera vez el proceso de enseñanza-aprendizaje desde la perspectiva de los aprendientes; por ello, se considera el análisis contrastivo (AC) el primer modelo de análisis centrado en la adquisición de la L2, aunque el acercamiento científico fue casi completamente teórico, pues apenas produjo resultados prácticos concretos (De Alba, 2009). En plena vigencia de la teoría lingüística estructuralista y del modelo de aprendizaje conductista, se consideraba el error como señal de no adquisición –por lo que debía ser penalizado– y de interferencia de los hábitos de la L1. A partir de la relación entre los errores de los HNN y la diferencia entre la L1 y la L2, los investigadores del AC trataron de localizar la fuente del error por medio de la comparación entre las dos lenguas: "In the comparison between native and foreign language lies the key to ease or difficulty in foreign language teaching" (Lado, 1971: 1 [1957]; véase como muestra Stockwell *et al.,* 1965). El objetivo de la comparación era identificar los rasgos fáciles y difíciles de la L2; la L1 podía ayudar a los HNN o interferir negativamente durante la producción de las estructuras gramaticales y léxicas. De esta manera, se presuponía que había una relación directamente proporcional entre la distancia lingüística y la dificultad del aprendizaje (Pastor Cesteros, 2004: 101).

Este método de análisis conllevó una serie de problemas o carencias que los profesores de lenguas detectaron pronto; así lo señala Corder (1967: 161):

> Teachers have not always been very impressed [by the contributions of the Contrastive Analysis Research] for the reason that their practical experience has usually already shown them where these difficulties lie and they have not felt that the contribution [of this research] has provided them with any significantly new information.

Además, los profesores percibían que dos lenguas más distantes no generaban más dificultades que dos cercanas, como el AC defendía –pese a que sí que existiera, y existe, un concepto de distancia interlingüística, incluso de tiempo de estudio, basado en las diferencias tipológicas de las lenguas–. En esta misma línea, no se demostró que por medio de este método se pudieran predecir todos los errores producidos en el aula, ni se logró probar que todos los errores predichos se materializaran.

Deben reconocerse, sin embargo, los logros del AC: inicia la investigación centrada en el aprendiente y su proceso de aprendizaje, y supone el punto de partida de los actuales estudios de ASL (Pastor Cesteros, 2004: 103); de hecho, el desarrollo de esta disciplina se vincula a la comprobación o negación de las propuestas del AC, con lo que es obligada su mención en cualquier recorrido por la evolución de las investigaciones sobre la adquisición y el aprendizaje de las L2, como también argumenta De Alba (2009). Así, conceptos que provienen de este modelo de análisi, como *interferencia*, *evitación*, *sobreuso* o *infrauso*, o el influyente factor de la lengua materna, que está más presente de lo que parece en el proceso de adquisición de una L2, así como la hipótesis del marcado diferencial, son hoy muy relevantes en la investigación de la interlengua.

Con respecto al desarrollo del AC en relación con el aprendizaje del español, la *Bibliografía de lingüística general y española (1964-1990)* (Báez, 1995), que abarca un periodo extenso de tiempo, muestra la existencia de un conjunto amplio de trabajos en los que se analizan temas de fonología, morfología y sintaxis del español y el inglés, o del español y el francés. Estos trabajos representan lo que se conoce como la *versión débil* del AC, que destaca el carácter explicativo –pero no predictivo– de la interferencia lingüística para entender las dificultades en el aprendizaje. En Penadés (1999: 9) se hace referencia también a la existencia de algunos trabajos, menos numerosos que los anteriores, que contrastan el español con otras lenguas, como el italiano, el rumano, el alemán, el portugués, el checo, el ruso, el japonés o el sueco. Asimismo, los números 51 y 52 de la revista *Carabela*, del año 2002, que están dedicados a la lingüística contrastiva, presentan trabajos muy completos con este enfoque.

Así, aunque no han dejado de realizarse estudios contrastivos –de hecho, a finales del siglo XX se hablaba de una revalorización de la Lingüística Contrastiva (Fernández González, 1995: 14-19)–, las críticas a este modelo de análisis conllevaron la aparición de un nuevo modelo, el análisis de errores, que supuso un gran avance en los estudios de adquisición al tener como objeto de estudio la lengua del aprendiente, y no ya la L1 ni la L2.

1.1.2. *El análisis de errores*

A finales de los años 60 y durante los años 70 del siglo XX, la investigación de la lengua del aprendiente se correspondía con el análisis de errores (AE), que resultó en una práctica muy extendida. Este método de análisis se erigió en el nuevo modelo de investigación de la adquisición de la LE, tras constatarse empíricamente la incapacidad del AC para alcanzar los objetivos propuestos por Lado (1971 [1957]) y tras producirse la caída de sus bases teóricas –representadas lingüísticamente por el estructuralismo y psicolingüísticamente por el conductismo–, provocada por la irrupción de los planteamientos teóricos de la lingüística chomskiana y de las teorías cognitivas del aprendizaje. En ese momento comienzan a interesar las producciones concretas de los HNN:

> el cambio metodológico es crucial: se pasa de las predicciones desde el plano de la abstracción –en el que hasta ahora se habían desarrollado las investigaciones en L2– al espacio real y concreto de las producciones de los discentes, con el objeto de obtener datos empíricos que favorezcan la explicación de los errores en el proceso de adquisición y aprendizaje de las lenguas extranjeras (De Alba, 2009: s/p).

Desde el AE se concibe la adquisición de una segunda lengua como un proceso cognitivo e interiorizado de formación de reglas de la L2 –lo que supone un traslado de la hipótesis innatista al respecto de la adquisición de la L1 al ámbito de la adquisición de segundas lenguas–, en el que se establece una línea de continuidad entre la L1 y la L2. El sistema lingüístico que se desarrolla a través de ese proceso se conoce como *interlengua*, y nace, precisamente, en el seno del AE (Pastor Cesteros, 2004: 105). Corder (1967) consideraba la lengua del aprendiente una especie de *dialecto idiosincrásico*, con sus propios rasgos, diferentes de la L1 y de la L2, en el que el error constituye un aspecto fundamental, pues comporta un gran valor para el estudio del proceso de aprendizaje e informa del nivel de adquisición; esto es, entre los investigadores del AE existe la opinión compartida de que los errores de los HNN, lejos de ser considerados negativos, son inevitables y necesarios en el desarrollo del aprendizaje, pues constituyen evidencias positivas de que el aprendizaje está teniendo lugar:

> A learner's errors, then, provide evidence of the system of the language that he is using (i.e. has learned) at a particular point in the course (and it must be repeated that he is using some system, although it is not yet the right system). They are significant in three ways. First to the teacher, in that they tell him, if he undertakes a systematic analysis, how far towards the goal the learner has progressed and consequently, what remains for him to learn. Second, they provide to the researcher evidence of how language is learned or acquired, what strategies or procedures the learner is employing in his discovery of the language. Thirdly (and in a sense this is their most important aspect) they are indispensable to the learner himself, because we can regard the making of errors as a device the learner uses in order to learn. It is a way the learner has of testing his hypothesis about the nature of the language he is learning (Corder, 1967: 169).

De esta manera, el error se convierte en objeto de estudio y es utilizado con fines didácticos. El tratamiento del error planteado por Corder (1971) se puede resumir en cuatro etapas (Ellis, 1994: 68-69): la recogida de errores a partir de la recopilación de muestras de lengua de aprendientes, su identificación, su descripción y su explicación.

> The first step in carrying out an EA was to collect a massive, specific, or incidental simple of learner language. The sample could consist of natural language use or be elicited either clinically or experimentally. It could also be collected cross-sectionally or longitudinally. The second stage involved identifying the errors in the sample. Corder distinguished errors of competence from mistakes in performance and argued that EA should investigate only errors. [...] The third stage consisted of description. Two types of descriptive taxonomies have been used: linguistic and surface strategy. The former provides an indication of the number and proportion of errors in either different levels of language (i.e. lexis, morphology, and syntax) or in specific grammatical categories (for example, articles, propositions, or word order). The latter classifies errors according to whether they involve omission, additions, misinformations, or misordering. The fourth stage involves an attempt to explain the errors psycholinguistically.

Así, en la primera etapa o recopilación de muestras, el análisis del error llevado a cabo por los investigadores del AE puede estar basado en datos procedentes de test de elicitación y datos de corpus preinformatizados[2]; puede centrarse en un solo individuo (estudios

[2] El término *corpus* en este periodo se refiere a una colección de muestras de lengua natural que ha sido compilada para un estudio lingüístico (Hunston, 2002: 2). En la actualidad, el término implica que el corpus se almacena y que se accede a él electrónicamente.

longitudinales, que involucran a HNN que son seguidos durante un periodo de tiempo) o en un grupo representativo de una población (estudios longitudinales o transversales, estos últimos basados en distintos niveles de competencia). No obstante, la capacidad de generalización de los resultados de estos estudios está vinculada al número de representantes y al número de muestras, aunque, como bien observa De Alba (2009), no existe ningún criterio normalizador en este ámbito.

En lo que se refiere a la segunda y tercera etapa en el tratamiento del error (identificación y descripción), Corder ya señaló que no se pueden considerar errores todos los fallos que comete el HNN, sentando las bases para diferenciar entre error (fallo sistemático, relacionado con la competencia) y falta (fallo asistemático, relacionado con la actuación); tras la identificación del error, se suele establecer una taxonomía, previa o posterior, para clasificar los errores hallados. En la cuarta etapa, en fin, se analizan las causas de los errores.

Aunque sin duda este proceso de análisis aporta información muy valiosa sobre el modo en el que se produce la adquisición de la L2, no son pocas sus limitaciones. Schachter y Celce-Murcia (1977: 441), en pleno apogeo del AE, expusieron sus reservas con respecto a este método. Como también señala Guo (2006), de entre todas sus limitaciones sobresalen tres por su vinculación directa con la caída del AE y la llegada de nuevos modelos de análisis de la lengua del aprendiente: el análisis de los errores aislados, la clasificación de los errores identificados y la adscripción de las causas a los errores. En relación con la primera, los investigadores extraían de los datos los errores de los HNN; una vez que eran identificados, los datos se descartaban, por lo que no existía la posibilidad de recuperarlos para confirmar resultados o reevaluar el contexto. En segundo lugar, está la dificultad para identificar los errores –pues no siempre es fácil determinar qué es un error y qué es una falta, ni si un determinado uso es verdaderamente una desviación de la L2 o no lo es– y la dificultad para clasificarlos, esto es, establecer con qué tipo de estructura se corresponde un determinado error, pues, como señala Penadés (2003), existen numerosas taxonomías en relación con la clasificación de los errores, lo que ha generado mucha controversia debido a la falta de criterios normalizadores. Este aspecto se refleja en dos procedimientos presentes en la investigación que exponemos aquí: el establecimiento de un concepto instrumental de *norma* con el que contrastar los datos –para así poder determinar qué es un error y qué no lo es– y, por otro lado, la fijación de unos parámetros en la clasificación de los datos, en general, y del error, en particular, una vez que los errores son identificados –y no antes–, para no influir en el análisis con los tipos de errores que previsiblemente nos podríamos encontrar. Por último, existe el problema de la adscripción de las causas a los errores. A diferencia del AC, en el AE las causas no se restringen a la trasferencia interlingüística (Hammarberg, 2009 [1973]); estas pueden ser numerosas entre las interlinguales y las intralinguales, pero es una práctica común entre los investigadores de AE hacer un análisis de errores aislados dentro de un alcance muy limitado y luego etiquetarlos como intralinguales o interlinguales (Guo, 2006: 10). Por otro lado, pero en relación con este punto, Dulay *et al*. (1982) manifiestan que uno de los principales problemas del AE es que en los estudios se mezclan la descripción del error y la explicación, esto es, qué es lo incorrecto en un determinado uso y por qué. Se podría evitar esta situación si –como señala De Alba (2009)– en cada taxonomía existiera una breve explicación, por cada uno de los apartados en los que se ha dividido la clasificación, de las reglas que han sido trasgredidas antes de abordar las explicaciones de estos conflictos.

Es evidente que el tratamiento de los errores aislados y la complejidad y la falta de sistematicidad para identificar y clasificar los errores son debilidades importantes de este modelo; pero no menos problemático es el hecho de tener como objeto de estudio únicamente los errores, lo cual no revierte en un conocimiento completo sobre la lengua producida por el HNN, como pronto reivindicaron Hammarberg y Enkvist en sendos trabajos: "The insufficiency of error analysis" (1973) y "Should we count errors or measure success?" (1973). Svartvik sugiere ya en ese mismo año (1973) que la expresión *análisis del error* debería ser reemplazada por la de *análisis de actuación*, aunque finalmente la que se estableció fue la de *estudios de interlengua*, en relación con el término *interlengua*, de Selinker (1972) (Hasselgard y Johansson, 2011: 35): "Although the study of errors is a natural starting-point, the final analysis should include linguistic performance as a whole, not just deviation" (Svartvik, 1973: 8)[3].

Ellis (1994: 67) se refiere a esta misma idea cuando años después señala: "A frequently mentioned limitation is that EA fails to provide a complete picture of learner language. We need to know what learners do correctly as well as what they do wrongly". Así pues, queda constatado que el objetivo del estudio de la lengua del aprendiente no se aplica solo a los errores, sino que debe representar el nivel de dominio del HNN.

Debido a los problemas metodológicos señalados, el AE fue gradualmente absorbido por un campo de estudio de la adquisición de la L2 más general, el conocido hoy como ASL (Guo, 2006: 3, 10-11). Pero, pese a estas limitaciones, el AE, orientado hacia la descripción de la interlengua, ha sido el modelo de investigación de la adquisición del español más productivo hasta ahora; no obstante, como señala Baralo (2004: 38 [1999]), la fase explicativa del análisis del error no se ha desarrollado totalmente. Algunos de los trabajos más exhaustivos han sido el de Vázquez (tesis doctoral defendida en 1987 y publicada en 1991) –centrado en los errores principalmente morfosintácticos producidos en la interlengua de estudiantes alemanes–, Fernández (tesis defendida en 1991 y parcialmente publicada en 1997) –que examina los errores en todo el sistema de la lengua en cuatro grupos de L1 diferentes, en tres estadios de evolución de su interlengua– y Santos Gargallo (tesis defendida en 1992 y publicada al año siguiente) –que examina los errores de la producción escrita en alumnos serbocroatas–. Disponemos también de numerosos estudios menos abarcadores de análisis de errores de HNN de español, como el de Penadés (1999), que incluyen memorias de maestría de alumnos de la Universidad de Alcalá. Estos se suman a otras tesinas, tesis doctorales y otros trabajos de investigación que utilizan esta metodología a finales de los años 90 y principios del siglo XXI; son buena prueba de ello, pues incluyen muchos trabajos de investigación centrados en el AE, las actas de numerosos congresos de ASELE (véanse las de 2005, 2006 o 2007, por poner algunos ejemplos), la base de datos teseo, de tesis doctorales realizadas en España (https://www.educacion.gob.es/teseo/irGestionarConsulta.do) y la Red Electrónica de Didáctica del Español como Lengua Extranjera del Ministerio (http://www.mecd.gob.es/redele/Biblioteca-Virtual/Presentacion.html). La evolución de estos estudios muestra que a partir de los análisis únicamente lingüísticos se ha ido pasando a los análisis que integran también los elementos discursivos, aunque los primeros

[3] La investigación de Linnarud, de 1986 –iniciada en el contexto del proyecto de estudios contrastivos sueco-inglés, dirigido por Svartvik (1973)–, es un análisis de la actuación léxica, no solo de los errores. Este estudio usa material textual combinado con material procedente de técnicas de elicitación.

en ningún momento han dejado de producirse. Un análisis de errores puede centrarse en cualquiera de las cuatro destrezas, aunque las dos productivas (expresión oral y escrita) son las más estudiadas, dado que en ellas es más fácil cuantificar datos (De Alba, 2009).

El tratamiento del error que se lleva a cabo en el AE –propuesto por Corder (1967, 1971)– constituye una contribución que sigue teniendo vigencia hasta ahora en los análisis de interlengua o de la actuación, ya estén basados en corpus o no lo estén, en tanto en cuanto los errores forman parte de la lengua de los HNN, es decir, contribuyen a determinar el nivel de dominio lingüístico en el que se encuentran los HNN. La investigación que desarrollamos aquí no pertenece al AE, pero sí que adopta y adapta la metodología del AE para analizar el error, como luego se verá.

1.1.3. *La ASL y el problema de los datos*

Tras varias décadas de desarrollo desde los años 70, la investigación de ASL se ha abordado desde múltiples perspectivas y enfoques teóricos y prácticos. Larsen-Freeman y Long (1994 [1991]) sostienen que el ámbito de la ASL es fundamentalmente la naturaleza del proceso de adquisición de la L2 y los factores que afectan a la lengua de los HNN[4], que constituye, al cabo, la principal fuente de datos para analizar dicho proceso de adquisición. Al respecto, la obtención de datos, cuya validez y fiabilidad determinan el éxito de la investigación, se presenta como una de las principales limitaciones de la investigación en ASL. La mayoría de los estudios apuesta por los datos experimentales (elicitación) y los procedentes de la introspección, y tiende a rechazar los datos de usos lingüísticos naturales, que suelen estar representados en los corpus. Granger (2002: 5-6) explica esta preferencia en la investigación de ASL refiriéndose a la dificultad que existe en los contextos no experimentales para controlar las variables que afectan a la producción de los HNN; además, por medio de estos procedimientos experimentales el investigador se asegura de que la estructura que desea investigar está presente en el material analizado, y de que lo que está presente lo está porque el HNN lo conoce. A esto hay que añadirle la falta de formación de los lingüistas aplicados en el uso de las metodologías informatizadas, que son las que permiten trabajar con datos naturales a gran escala, como señala Tono (2003: 806).

Como es difícil someter a una gran cantidad de informantes a la experimentación, la investigación de la ASL tiende a emplear una base empírica relativamente estrecha, con el foco puesto en la lengua de un número muy limitado de individuos, lo que provoca cuestionamientos acerca de la generalización de los resultados (Granger, 2002: 6) –como en el estudio de Lardiere (1998), que usa los datos de un solo aprendiente de inglés recopilados durante varios años–. Esto, por consiguiente, demuestra que la lengua colectiva del aprendiente no es foco de interés de la investigación actual de la ASL (Guo, 2006: 3): "When the learner's *output* is considered, the focus of the research is rather more on the *output* of individual learners than on the *output* of a group of learners with the same background".

[4] Trabajos como los de Corder (1969), Selinker (1972), Schumann (1976) y Krashen (1977) son representativos del nacimiento de la ASL; en ellos se separa la enseñanza del aprendizaje y "se sientan las bases de lo que luego se va a consolidar como acercamientos psicolingüísticos, lingüísticos y sociolingüísticos al estudio de la adquisición" (Muñoz Liceras, 2009).

Esto resulta llamativo cuando precisamente las dificultades típicas de un grupo concreto son las que deben atenderse en la enseñanza de lenguas y, por lo tanto, considerarse en las investigaciones de la interlengua.

Por ello, Granger (1998, 2015), Guo (2006: 11) o Lozano y Mendikoetxea (2013: 1-3), entre otros, reivindican el uso de los corpus de aprendientes informatizados como una fuente de datos naturales muy valiosa: "There is clearly a need for more, and better quality, data and this is particularly acute in the case of natural language data […] learner corpora are a valuable addition to current SLA data sources" (Granger, 1998: 5). Lozano y Mendikoetxea (2013: 1-2) argumentan con acierto el hecho de que los estudios sobre la adquisición de la L1 se han servido del uso de corpus extensos informatizados durante los últimos 25 años, como el Child Language Data Exchange System o CHILDES (http://childes.psy.cmu.edu/), que abarca tanto la dimensión monolingüe del lenguaje como la bilingüe, la no nativa y la patológica, y que ha sido la fuente de datos de más de 3200 trabajos, lo que ha supuesto un gran paso en el conocimiento sobre la manera en que los niños adquieren y desarrollan su L1 (MacWhinney, 2000). Los corpus en L2 son todavía escasos, con lo que la investigación se ha podido beneficiar poco hasta ahora de esta fuente de datos naturales a gran escala que tiene mucho que aportar a nuestro conocimiento sobre el modo en que se desarrolla la lengua del aprendiente.

En general, la contribución de la investigación de corpus de aprendientes informatizados (CAI) ha sido más substancial en la descripción que en la interpretación de los datos referidos a la ASL (Granger, 2004; Lozano y Mendikoetxea, 2013: 3); más centrada en un acercamiento pedagógico a la ASL, con pocas referencias a los debates, hipótesis y teorías actuales sobre la ASL y sus implicaciones en el desarrollo de la lengua del aprendiente (Myles, 2005) –como se verá a continuación, nuestro trabajo se enfoca hacia esta nueva dirección–; en su lugar, los estudios de corpus, al igual que los estudios experimentales, han sido útiles en el intento de construir hipótesis en la investigación en la ASL.

1.2. La investigación de corpus de aprendientes

Los comienzos de la LC se remontan a los años 60 del siglo XX, cuando se compilan los primeros corpus que permiten mejorar las descripciones del inglés. Como ya se ha señalado, en el ámbito específico de la LC el significado de *corpus* es mucho más restringido que el que se utiliza en la lengua general o en el periodo anterior al nacimiento de esta rama de la lingüística, dado que se concibe necesariamente en soporte electrónico (Sinclair, 2005: 16).

Ya desde los años 70 es común el uso de corpus en la investigación sobre la adquisición de la L1. El corpus más extenso es el ya mencionado CHILDES, que cuenta con 44 millones de palabras en más de 30 lenguas diferentes, y constituye un referente en el estudio de la adquisición de la L1 y del bilingüismo (Lozano y Mendikoetxea, 2013: 3). No obstante, los corpus de aprendientes informatizados no aparecieron hasta comienzos de los 90, cuando la tecnología y el análisis de corpus nativos se encontraban relativamente desarrollados, tras originarse el interés, a partir de finales de los 80, por la creación de recursos de corpus útiles de acuerdo con las necesidades de los aprendientes de lenguas extranjeras (Tribble, 2011: 85).