La Investigación de Corpus de Aprendientes y el desarrollo de los estudios de la interlengua del español

ANNA SÁNCHEZ RUFAT
Universidad de Extremadura

Resumen

Este artículo recoge los avances y cambios que ha habido en la investigación de la lengua del aprendiente, o interlengua, desde los orígenes de esta práctica en los años 40 con el Análisis Contrastivo hasta el actual estudio de corpus informatizados de aprendientes. A través de un recorrido por los diferentes modelos de análisis de datos –que incluye el Análisis de Errores, y su ulterior absorción por el campo de estudio más general conocido como Adquisición de Segundas Lenguas, y el Análisis de Errores Asistido por Ordenador–, se pone de manifiesto que el aspecto colectivo de la lengua del aprendiente de español ha recibido poca atención en la investigación actual sobre la Adquisición de Segundas Lenguas. La Investigación de Corpus de Aprendientes puede suplir estas carencias, y sumar otras ventajas que son identificadas en este estudio.

Palabras clave: *interlengua, lingüística de corpus, corpus de aprendientes, análisis contrastivo de la interlengua, análisis de errores*

Abstract

This article refers to the progress and changes that have been made in language learner research, from the origins of this practice in the 40s with the Contrastive Analysis to the current Contrastive Interlanguage Analysis based on corpus. This review of the different models of data analysis –which includes the Error Analysis and its subsequent absorption by the more general field of study known as Second Language Acquisition, as well as the Computer-aided Error Analysis– reveals that the collective aspect of the learner's language (of Spanish) has received little attention in the current research on Second Language Acquisition. Computer Learner Corpora can fill these gaps, and add other advantages that are identified in this study.

Key words: *Interlanguage, Corpus Linguistics, Computer Learner Corpus, Contrastive Interlanguage Analysis, Error Analysis*

1. Los orígenes de los estudios de corpus de aprendientes y la lingüística de corpus

Aprender una lengua extranjera es un proceso lento y costoso, que en raras ocasiones culmina en un dominio completo de la lengua meta. Los problemas pueden estar relacionados con la interferencia de la lengua materna, con los rasgos de la lengua meta o con el propio proceso de aprendizaje. Revelar los rasgos de la lengua del aprendiente, o de la interlengua[1], se ha convertido en un medio importante de medir las diferencias entre la interlengua y la actuación del hablante nativo, lo que puede llevar potencialmente a una mejora en la enseñanza de la lengua y en el conocimiento del proceso de aprendizaje de la lengua (Hasselgard y Johansson, 2011: 33).

La investigación basada en corpus de aprendientes es una rama de estudio sobre la lengua del aprendiente relativamente nueva, cuyo desarrollo supuso un paso significativo en los estudios de interlengua. En este ámbito específico, *corpus* tiene un significado restringido: "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, 2005: 16).

Esta rama de estudios nace en la década de los noventa de la mano de Sylviane Granger, quien –tras fundar en 1990 el Centro de Lingüística de Corpus del Inglés (CECL[2]) de la Universidad Católica de Lovaina– inicia un exitoso proyecto de compilación del primer corpus internacional de aprendientes de inglés denominado International Corpus of Learner English (ICLE, Granger *et al*., 2002; 2009 [segunda versión]), que ha inspirado

[1] *Interlengua* es el término introducido por Selinker (1972) para referirse al sistema estructurado que construye el aprendiente de la LE en un estadio determinado del desarrollo del aprendizaje. Es un sistema lingüístico dinámico que no se concibe como defectuoso. Este concepto es también denominado *sistema aproximado* (Nemser, 1971) o *competencia transitoria* (Corder, 1967); de aquí en adelante utilizaremos los términos *interlengua* y *lengua de aprendientes* indistintamente para aludir al sistema lingüístico no nativo, por ser los más extendidos para referirse al sistema lingüístico no nativo.

[2] Para una información completa sobre el CECL, véase www.uclouvain.be/cecl.html (10-06-14).

trabajos similares en muchos otros países, como es el caso del CEDEL2 (Lozano, 2009; Lozano y Mendikoetxea, 2013) en España.

Estos corpus surgen más de treinta años después de que comenzaran a compilarse los primeros corpus de hablantes nativos. El objetivo principal de los lingüistas y los lexicógrafos era entonces construir una representación de la producción escrita y hablada de los hablantes nativos de una lengua, lo que supuso el inicio del desarrollo de los corpus informatizados. En rigor, puede decirse que la revolución de los corpus en lingüística –*revolución* en el sentido de que ha originado toda una relación de nuevas preguntas de investigación– comenzó con la finalización y la distribución del Brown Corpus, en 1964, publicado por Kučera y Francis, de la Universidad de Brown, con el título *A Standard Sample of Present-Day Edited American English, for Use with Digital Computers* (Leech, 2011: 10). Poco después, Kučera y Francis (1967) lo usaron para crear la primera lista de frecuencias del inglés basada en datos de un corpus; en 1982 publicaron listas de frecuencias lematizadas, basadas en la versión del corpus anotada con las categorías de palabras (Leech, 2011: 10). En ese mismo año, los lingüistas escandinavos Hofland y Johansson compararon el trabajo de Brown con el que ellos acababan de concluir, el Corpus of British English Lancaster-Oslo/Bergen (LOB), y publicaron el trabajo titulado *Word Frequencies in British and American English*. Por primera vez, listas de palabras frecuentes –del inglés americano y el británico– con información gramatical derivadas automáticamente de un corpus informatizado estaban disponibles para el investigador y para el profesor de lenguas. Pese a su innovación, estas técnicas ya habían sido empleadas en un trabajo más revolucionario –*The OSTI Report*– que implicaba un cambio en la concepción de la lengua, sobre todo en la relación entre la gramática y el léxico. Fue dirigido por Sinclair desde la Universidad de Birmingham y concluido en 1970 pero no llegó a ver la luz hasta el año 2004[3]. Este trabajo usaba un pequeño corpus de inglés hablado de 135000 palabras, que en los años ochenta y noventa fue ampliado bajo la dirección de este mismo lingüista en lo que se denominó *Birmingham Collection of English Texts*, conocido después como el *Bank of English* (1997).

[3] Sobre los orígenes, avances y progresos en la descripción de la lengua basada en los análisis de corpus, véase Sinclair *et al*. (2005).

Estos trabajos se encuentran en la base de algunos de los estudios más recientes de la interlengua del español, que analizan aspectos de la interfaz léxico-sintaxis (Prieto González *et al*, 2009; Alonso Ramos *et al.* 2010a, 2010b; Prieto González *et al*. 2011; Orol González y Alonso Ramos, 2013; Wanner *et al*. 2013; Sánchez Rufat, 2014; Sánchez Rufat, *en prensa*), una de las cuestiones más debatidas en este momento en la investigación de ASL al ser considerada una potencial fuente de déficits y carencias en el desarrollo de la interlengua (Lozano y Mendikoetxea, 2013: 6). Estos estudios que, como se ha señalado *supra*, parten de la interrelación entre los niveles gramatical y léxico y analizan el vínculo entre la combinatoria léxica de una palabra y la estructura de su significado, son herederos, en cierta medida, de los trabajos de Sinclair (1966, 1970) –surgidos a raíz de las ideas desarrolladas por Firth (1957)– y de Halliday (1961, 1966), su colega entonces en Edimburgo, quien declaró (1966: 273-277) que el nivel del léxico (incluyendo la colocación) merecía constituir un nivel de descripciónlingüística diferenciado de los demás, interrelacionado en un continuo con el nivel gramatical; los corpus de hablantes nativos de inglés daban muestra de ello[4]. En este enfoque, las palabras y la gramática están ineluctablemente unidas en lo que se conoce como *lexicogrammar* (Halliday, 1985; Sinclair, 1991) o *lexical grammar*, lo que supuso un impacto en el campo de la lexicografía: "Corpus data forced lexicographers to shift away from a focus upon regularities which were easily attested by a few examples towards what Hanks (2009: 216) describes as 'idiosyncratic conventions that are associated with each word'" (McEnery y Hardie, 2012: 80). En este sentido, puede decirse que

> *the corpus revolution has introduced a new theoretical perspective on linguistic structuring: one in bold contrast to the mainstream paradigm of Chomsky (e.g. Chomsky 1965: 84-88) whereby grammar and lexicon are two clearly distinct components. It also challenges a tradition long established in language study, whereby grammars and dictionaries provide distinct kinds of information about a language, and are published in separate covers (Leech, 2011: 12).*

[4] Las principales contribuciones sistemáticas de la Lingüística de Corpus a la descripción del léxico y la gramática se hicieron dentro de la Lingüística de Corpus del inglés (McEnery y Hardie, 2012: 72), donde se desarrollaron y refinaron conceptos como los de colocación y anotación de corpus.

En relación con esto, han sido publicadas recientemente obras lexicográficas que contienen información de cierta naturaleza lingüística que no aparece en las tipologías de diccionarios. Nos referimos a los diccionarios combinatorios *Redes* (2004) y *Práctico* (2006) –basados en un corpus de 250 millones de palabras–, que supusieron una innovación en el terreno lexicográfico al proporcionar la combinatoria de cada palabra basada en la restricción semántica que los predicados ejercen sobre sus argumentos, al tiempo que incluyen información sobre la frecuencia de uso de las combinaciones. Estos diccionarios de restricciones léxicas no tienen equivalentes directos en ningún diccionario del español o de cualquier otro idioma (Bosque, 2005: LXXXVI); y unidos a los diccionarios de colocaciones –basados también en análisis de corpus– ponen de manifiesto la estrecha relación que existe entre el trabajo lexicográfico, el lexicológico y el gramatical, como también puede verse en el recientemente publicado diccionario de frecuencias del inglés americano (Davies y Gardner, 2010), basado en un corpus de 385 millones palabras, que presenta, junto a la frecuencia individual de cada palabra, una lista de colocaciones comunes para cada palabra (estas cifras muestran el aumento del tamaño de los corpus utilizados en los últimos cuarenta años).

En suma, se puede decir que la lingüística de corpus se ha convertido en los últimos veinte años en un componente indispensable del aparato metodológico de la lingüística, en general. Esto es, en rigor, se trata más bien de una metodología que de una teoría del lenguaje; una metodología que ha facilitado nuevos accesos para el estudio de la lengua –su naturaleza y la manera en la que la procesamos– y nuevas maneras de unir teoría y datos (Altenberg, 2011: XIII). Así lo señalaba Lewis (2000: 126) "In recent years, since the widespread availability of large computer-based corpora – collections of natural written and spoken text– which have been statistically analysed, we have better descriptions of English available to us than ever before".

Así pues, desde hace ya varios años la lingüística de corpus y otras ramas de la lingüística tienden a converger (McEnery y Hardy, 2012: 226); es decir, la metodología de corpus forma parte de la práctica diaria de lingüistas funcionalistas, sociolingüistas, analistas del discurso o de estudiosos de la lengua de los aprendientes, entre otros. En relación con ello, Granger (1998a: XXI) resume de la siguiente manera esta última rama de estudios centrada en la lengua de aprendientes y su análisis a través de los corpus: "With roots both

in corpus linguistics and second language acquisition studies (SLA), it uses methods and tools of corpus linguistics to gain better insights into authentic learner language". A los orígenes de los estudios de Corpus de Aprendientes Informatizados (CAI) y la lingüística de corpus nos acabamos de referir. En cuanto a su relación con la Adquisición de Segundas Lenguas (ASL), dado que han sido varios los procedimientos en la investigación del proceso de adquisición anteriores a los estudios de CAI de los que estos se nutren –como el Análisis Contrastivo (AC), el Análisis de Errores (AE) y el Análisis de la Interlengua o de la actuación (AI)–, se incluye a continuación una revisión de estos procedimientos (2.1 y 2.2). Después, se señala la relación entre estos y la ASL, y se plantea la conexión entre esta disciplina y la investigación de CAI (2.3). Por último, el apartado final está dedicado a la investigación de CAI (3).

2. Estudios sobre la lengua del aprendiente antes de los corpus informatizados

2.1. El Análisis Contrastivo

La lengua del aprendiente fue investigada antes de la aparición de los corpus de aprendientes; se pueden datar los orígenes del análisis de este objeto a finales de los años sesenta y en los años setenta, con el AE. No obstante, antes del periodo del AE, entre los años cuarenta y sesenta, surgen los primeros estudios de adquisición –a medio camino entre los estudios lingüísticos y psicolingüísticos– en un momento en el que se está produciendo el debate sobre la teoría del aprendizaje lingüístico y el consiguiente enfrentamiento entre conductistas y cognitivistas (Pastor Cesteros, 2005: 369). Estos estudios orientan por primera vez el proceso de enseñanza-aprendizaje desde la perspectiva de los aprendientes; por ello, se considera el AC como el primer modelo de análisis centrado en la adquisición de la L2, aunque el acercamiento científico fue prácticamente teórico en su totalidad, 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 que el error debía ser penalizado, pues es señal de no adquisición y de interferencia de los hábitos de la L1. Desde que los investigadores del

AC establecieron una relación entre los errores de los aprendientes y la diferencia entre la L1 del aprendiente y su L2, 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, 1957: 1); véase, como muestra del AC aplicado al aprendizaje de ELE, *The grammatical structures of English and Spanish*, de Stockwell *et al.* (1965), una de las obras más representativas de este modelo de análisis. Por consiguiente, el objetivo de la comparación era identificar los rasgos fáciles y difíciles de la L2; la L1 podía ayudar a los aprendientes 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: a más diferencia mayor dificultad. 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.*

Los profesores percibían que dos lenguas más distantes no generaban más dificultades que dos cercanas, como el AC defendía. Además, no se demostró que por medio de este método se pudieran predecir todos los errores producidos en el aula; asimismo, no se logró probar que todos los errores que se predecían se materializaban.

No obstante, conviene reconocer 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, su desarrollo se vincula a la comprobación o negación de las propuestas del AC. Por todo ello, este tipo de análisis debe aparecer en cualquier recorrido sobre la evolución en la metodología de investigación del fenómeno de la adquisición y aprendizaje de las L2, como también argumenta De Alba (2009).

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 el español y el francés. 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 la L1 ni la L2, como observa Guo (2006: 2-3) "it was a significant advance when EA [error analysis] researchers to have placed the learner language (rather than L1 and L2) under examination"

2.2. El Análisis de errores

A finales de los años sesenta y durante los años setenta, la investigación de la lengua del aprendiente, conocida entonces como el AE, resultó en una actividad muy popular. Este método de análisis se erigió como el nuevo modelo de investigación de la adquisición de la lengua meta, tras constatarse empíricamente la incapacidad del AC para alcanzar los objetivos propuestos por Lado (1957) y tras 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 y mentalistas del aprendizaje. Como se ha referido anteriormente, lo que ahora comienza a interesar a los estudiosos son las producciones concretas de los aprendientes:

> *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, 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, concepto al que ya nos hemos referido (al comienzo de este capítulo[5]), y que 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 informa del proceso de aprendizaje y del nivel de adquisición; esto es, entre los investigadores del AE existe la opinión compartida de que los errores de los aprendientes, 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).*

[5] Véase n. 1.

De esta manera, el error se convierte en objeto de estudio y es utilizado con fines didácticos.

El tratamiento del error planteado por Corder se puede resumir en cuatro etapas (Ellis, 1994: 68-69), a saber: 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 pues, en cuanto a la primera etapa –la 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 (en Mairal Usón *et al*., 2010: 55, se enumeran algunos de estos procedimientos) y datos de corpus preinformatizado[6]; puede centrarse en un solo individuo (estudios longitudinales, que involucran a aprendientes 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 estos estudios

[6] 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). Esta definición está relacionada con la información de *corpus* obtenida del DRAE (2001): "conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación". En la actualidad, el término implica que el corpus se almacena y se accede a él electrónicamente.

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 aprendiente, sentando las bases para diferenciar 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 del error para clasificar los errores hallados, aunque esta también puede realizarse con anterioridad a la identificación o con posterioridad –se identifican los errores y posteriormente se clasifican–.

En la cuarta etapa, se analizan las causas de los errores, esto es, por qué se transgreden las normas.

Aunque sin duda este proceso de análisis aporta información muy valiosa sobre el modo en el que se produce la adquisición de segundas lenguas (Baralo, 2004; Guo, 2006, entre muchos otros), no son pocas las limitaciones de este modelo. Schachter y Celce-Murcia (1977: 41), en pleno apogeo del AE, expusieron sus reservas con respecto a este método de análisis, que evidencian las limitaciones del método a la hora de describir el proceso de adquisición de la L2. A nuestro juicio, de entre todas las 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 primer lugar, el error se analiza aislado; los investigadores extraían de los datos los errores de los aprendientes; 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 lengua meta o no lo es– y la dificultad para clasificar los errores, 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[7]. Por último, otro de los déficits asociados al AE es 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, 1973: 29); 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 que está 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[8].

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 aprendiente, 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

[7] Estos aspectos resultan fundamentales para realizar cualquier estudio de interlengua, pues conducen a reflexionar sobre la norma con la que contrastar los datos (véase a este respecto Sánchez Rufat y Jiménez Calderón, 2013) –para así poder fijar qué es un error y qué no lo es– y, por otro lado, a establecer unas pautas y unos parámetros en la categorización o catalogación de los datos, una vez que los errores son identificados, y no antes, para no influir en el análisis con los tipos de errores que previsiblemente se podrían encontrar.

[8] Véase Sánchez Rufat (*en prensa*) para una aplicación de esta cuestión metodológica al análisis del verbo *dar* en un corpus de aprendientes de español de nivel avanzado (CEDEL2).

(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)[9].

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 aprendiente.

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 Adquisición de Segundas Lenguas (Guo, 2006: 3, 10-11).

Pese a estas limitaciones, el AE, orientado hacia la descripción de la interlengua, ha sido el modelo de investigación de la adquisición de ELE más productivo hasta ahora, aunque, como señala Baralo (2004: 38), la fase explicativa del análisis del error no se haya desarrollado totalmente. Algunos de los trabajos más exhaustivos han sido los 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) –que examina los errores de la producción escrita en alumnos serbocroatas–. Disponemos también de numerosos estudios menos abarcadores de análisis de errores de aprendientes de español, como los de Penadés *et al*. (1999)[10], que incluyen tres memorias de maestría de alumnos de la Universidad de Alcalá. Estos se suman a otras

[9] 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 pruebas de elicitación.

[10] Para una idea más completa sobre la situación de la Lingüística Contrastiva, Análisis de Errores e interlengua del español en el último cuarto del siglo XX, véase el anexo al número 43 de la revista *Carabela*, donde se incluyen las referencias bibliográficas extraídas de publicaciones periódicas entre 1983 y 1997.

tesinas, tesis doctorales y otros trabajos de investigación que utilizan esta metodología a finales de los años 90 y durante los años transcurridos en el siglo XXI –las actas de los últimos congresos de ASELE (véanse las de 2005, 2006 o 2007, por poner algunos ejemplos), la base de datos Teseo[11] de tesis doctorales producidas en España y la red de didáctica del español como lengua extranjera del Ministerio[12] son buena prueba de ello, pues incluyen muchos trabajos de investigación centrados en el AE–. La evolución de estos estudios muestra que a partir de los análisis únicamente lingüísticos se ha ido dando paso a los análisis que integran también los elementos discursivos, aunque los primeros en ningún momento han dejado de producirse. Un análisis de errores puede orientarse hacia una subcompetencia o hacia todas las que constituyen la competencia comunicativa, y 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)– 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 aprendientes, es decir, contribuyen a determinar el nivel de dominio lingüístico en el que se encuentran los hablantes no nativos. Así pues, las investigaciones actuales de la interlengua no pertenecen al AE pero sí que adoptan y adaptan la metodología del AE para analizar el error.

2.3. Adquisición de segundas lenguas. Del prejuicio sobre los corpus de aprendientes a su apreciación en el análisis de la interlengua

Tras varias décadas de desarrollo desde finales de los años sesenta[13], la investigación de ASL "has become a rather amorphous field of study with

[11] Para hacer una consulta de esta base de datos, véase <https://www.educacion.gob.es/teseo/irGestionarConsulta.do> (12-06-14).

[12] Véase la página del Ministerio de Educación http://www.mecd.gob.es/redele/Biblioteca-Virtual/Presentacion.html

[13] 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

elastic boundaries" (Ellis, 1994: 15). Larsen-Freeman y Long (1994) sostienen que el ámbito de ASL es fundamentalmente la naturaleza del proceso de adquisición de la lengua y los factores que afectan a la lengua de los aprendientes. Esto es, el principal objetivo de ASL es construir modelos de representaciones mentales subyacentes –o de la interlengua– de aprendientes en un estadio particular en el proceso de adquisición de la L2 y de las restricciones que limitan la producción de la L2 (Lozano y Mendikoetxea, 2013: 1).

La principal fuente de datos para analizar este proceso de adquisición es la lengua producida por los aprendientes, ya sea espontáneamente o a través de los procedimientos de elicitación o de introspección (juicios de gramaticalidad, tareas de comprensión, etcétera). El éxito de la investigación en ASL depende de la validez y fiabilidad de estos medios de obtención de datos (Lozano y Mendikoetxea, 2013: 1-2). Precisamente, una de las principales limitaciones de la actual investigación en ASL está relacionada con la recopilación de datos. La mayoría de los estudios favorece los datos experimentales 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 al output de los aprendientes. A esto hay que añadirle la falta de formación de los lingüistas aplicados en el uso de las metodologías informatizadas 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)

Es cierto que existen razones para valorar las técnicas de elicitación, pues por medio de estos procedimientos no solo resulta más sencillo considerar las variables que afectan a la producción del aprendiente, sino que el investigador

---

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).

se asegura de que la estructura que desea investigar está presente en el material analizado, y que lo que está presente lo está porque el aprendiente lo conoce.

No obstante, como señala Granger, pese a estos beneficios, uno de los problemas principales es que la base empírica de la investigación en la ASL es muy estrecha, lo que pone en tela de juicio la generalización de los resultados, 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: "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" (Guo, 2006: 3). 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 deben interesar en las investigaciones de la interlengua.

Por ello, Granger (1998b: 5), Guo (2006: 11) o Lozano y Mendikoetxea (2013: 1-3), entre otros muchos, 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, 1998b: 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[14] (MacWhinney, 2000), 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 la que los niños adquieren y desarrollan su L1. Estos corpus en L2 son todavía escasos –aunque la situación está cambiando–. Consecuentemente, la investigación de la L2 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.

[14] CHILDES abarca tanto la dimensión monolingüe del lenguaje como la bilingüe, la no nativa y la patológica.

En general, la contribución de la investigación de corpus de aprendientes ha sido más substancial en la descripción que en la interpretación de los datos referidos a la ASL (Granger, 2004)15, 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, citado en Lozano y Mendikoetxea, 2013: 1-2) –en su lugar, los estudios de corpus, al igual que los estudios experimentales, han servido en el intento de construir hipótesis en la investigación en la ASL–. No obstante, esta situación está comenzando a dar un giro gracias a recientes estudios que buscan dar cuenta de la estructura de la interlengua probando hipótesis en torno al papel de las interfaces –la sintaxis-discurso (véase Lozano, 2009b; Lozano y Mendikoetxea 2008 y 2010) y la léxico-sintaxis (ya nos hemos referido a algunos de ellos en el apartado 1)–, "one of the much debated issues in second language research […], as a potential source of observed deficits in the development of learners' interlanguage grammars" (Lozano y Mendikoetxea, 2013: 6).

Así pues, queda constatado, como se ha señalado en el comienzo de este trabajo, que el campo de investigación lingüística conocido como investigación de Corpus de Aprendientes Informatizados ha surgido como resultado de la confluencia de dos campos hasta ahora discrepantes: Lingüística de Corpus y ASL.

En suma, se ha visto a lo largo de este recorrido por los diferentes modelos de análisis de la lengua del aprendiente que el AE no puede proporcionar una descripción completa de la lengua del aprendiente (recuérdese la reivindicación de Svartvik [1973: 8]), aunque sí intentó presentar el panorama de los errores de los aprendientes con claros propósitos pedagógicos; por otro lado, se ha puesto de manifiesto que el aspecto colectivo de la lengua del aprendiente ha recibido poca atención en la investigación actual sobre la ASL. Los corpus de aprendientes informatizados pueden suplir estas carencias, y sumar otras ventajas, como sugieren Leech (1998: xix), Nesselhauf (2005: 43) o Guo (2006: 12), entre otros. Estas se presentan en el siguiente apartado.

---

[15] Para un compendio de publicaciones recientes –basadas en corpus– más descriptivas que interpretativas, véase Lozano y Medikoetxea (2013: 3).

3. La investigación de corpus de aprendientes

Los comienzos de la LC se remontan a los años sesenta, como se ha referido anteriormente, cuando se compilan los primeros corpus que permiten mejorar las descripciones del inglés. 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, en los años sesenta, dado que se concibe necesariamente en soporte electrónico[16].

En los últimos 20 años se han confeccionado grandes corpus, como son, en el caso del inglés, el British National Corpus (BNC) –con 100 millones de palabras– y el Corpus of Contemporary American English (COCA) –con 410 millones de palabras–; y, en el caso del español, el Corpus del Español –con 100 millones de palabras– y el Corpus de la Real Academia Española (CREA) –con 154 millones de palabras–. La RAE ha puesto en marcha, con la colaboración de la Asociación de Academias de la Lengua Española (ASALE), el Corpus del Español del Siglo XXI (CORPES XXI). Este corpus de referencia y herramienta de investigación se presentó en el VI Congreso Internacional de la Lengua Española de Panamá, y reunirá textos reales de todos los países hispanohablantes, al igual que el CREA y el Corpus Diacrónico del Español (CORDE), pero con muestras de lengua recientes –desde el año 2001 hasta el 2012– que nos permitirán obtener mejores descripciones sobre el uso que se hace de la lengua durante el siglo XXI.

Ya se ha señalado que el uso de corpus en la investigación sobre la adquisición de la L1 tampoco es nuevo; desde los años 70 es una práctica común en los estudios de la lengua del niño. 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)

Ahora bien, los corpus de aprendientes informatizados no aparecieron hasta comienzos de los 90[17], cuando la tecnología y el análisis de corpus

[16] Véase n. 6 al respecto y la definición de *corpus* de Sinclair [2005: 16] recogida al comienzo del trabajo.

[17] Dentro del proyecto danés conocido como Project in foreign language pedagogy (PIF) se llevaron a cabo determinados estudios que se publicaron en el libro *Learner Language and Language Learning* (Færch *et al*., 1984, citado en Hasselgard y

nativos se encontraban relativamente desarrollados[18] –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).

En relación lógica con la ya expuesta definición de corpus de Sinclair (2005: 16), se entiende por *corpus de aprendientes*

> *Electronic collections of authentic[19] FL/SL textual data according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and are documented as to their origin of provenance (Granger, 2002: 7).*

En esta definición de Granger se sintetizan las características fundamentales de un corpus de aprendientes: es una recopilación de textos en soporte electrónico que requiere un diseño particular (de acuerdo con unos criterios estándar que son definidos por Sinclair [2005]). A los principales objetivos de este tipo de corpus se refiere la misma autora (Granger, 2002: 4): dispone de unas herramientas y métodos procedentes de la LC que ayudan a aportar descripciones mejoradas de la lengua del aprendiente que pueden ser utilizadas en la investigación de la adquisición y en la enseñanza de las L2:

> *Using the main principles, tools and methods from corpus linguistics, it aims to provide improved descriptions of learner language which can be used, for*

Johansson, 2011: 34), en el que se discuten muchos aspectos de la lengua del aprendiente a partir de un corpus de aprendientes único tanto en el tamaño como en el tipo. El corpus escrito contaba con 100000 palabras y el oral con 250000 palabras. Færch presentó en 1979 planes para informatizarlo, pero su muerte en 1987 impidió que el desarrollo del proyecto PIF fuera más allá y que se conformara el primer corpus de aprendientes informatizado (Hasselgard y Johansson, 2011: 34).

[18] El University College de Londres, la Universidad de Lancaster, la de Birmingham y la de Notthingham, en el Reino Unido; y las de Northern Arizona y Brigham Young, en EE.UU., han tenido mucho que ver en este desarrollo, mediante las aportaciones de lingüistas como Randolph Quirk, Geoffrey Leech, John Sinclair, Ron Carter, Mike McCarthy, Douglas Biber o Mark Davies.

[19] La autora especifica más adelante (Granger, 2002: 8) lo que se entiende por datos auténticos: los datos que resultan de actividades reales de clase.

> *a wide range of purposes in foreign/second language acquisition research and also to improve foreign language teaching.*

Así pues, surgen en esta época los corpus de aprendientes porque, por un lado, la LC ya está lo suficientemente consolidada desde un punto de vista tecnológico-metodológico, como ya se ha señalado, y, por otro lado, por las necesidades planteadas en el ámbito de la ASL que desembocaron en la confluencia de ambas ramas, por lo que los análisis de la interlengua basados en extensos corpus disfrutaron a partir de entonces de una buena acogida en este campo de la ASL. A estas necesidades nos referimos a continuación.

En primer lugar, los estudios basados en el análisis del comportamiento de los HN no informan de la dificultad de las estructuras específicas que deben ser enseñadas, ni de su proceso de aprendizaje; por ello, Granger (1998b: 7) reprueba que los materiales de enseñanza de inglés como lengua extranjera se diseñen "with a very fuzzy, intuitive, non-corpus-based view of the needs of an archetypal learner". Es imprescindible conocer las verdaderas dificultades de los aprendientes en el aprendizaje de la L2 para que puedan ser trabajadas adecuadamente en los materiales didácticos, y estas pueden ser fácilmente identificadas en los corpus de aprendientes. A este respecto, Leech (2001: 339) defiende que "corpus based interlanguage analysis enables us to identify areas of difficulty which are not derivable from Ns corpora alone". Estas ideas son recogidas por Nesselhauf (2004: 125-126) en la siguiente afirmación, en la que destaca para el desarrollo de la enseñanza de una L2 la importancia de disponer tanto de corpus nativos para conocer lo que estos verdaderamente utilizan como la de contar con corpus de aprendientes para averiguar sus dificultades:

> *Hardly anyone will doubt any longer that native speaker corpora are indeed useful for the improvement of language teaching. They are useful mainly because the can reveal –better that native speaker intuition– what native speakers of the language in question typically write or say (either in general or in a situation/in a certain text type). For language teaching, however, it is not only essential to know what native speakers typically say, but also what the typical difficulties of the learners of a certain language, or rather of certain groups of learners of this language, are.*

En segundo lugar, se hacen necesarios los corpus de aprendientes informatizados por las ventajas que ofrecen con respecto a los procedimientos adoptados normalmente en el AE tradicional (Altenberg y Granger, 2001: 189; Nesselhauf, 2005: 41), como son los corpus textuales preinformatizados y las pruebas de elicitación –de donde se obtienen datos seminaturales, ya que las tareas están diseñadas para controlar la lengua que los aprendientes deben producir–. Además, 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:

> *Unless they are very large, learner corpora with a great number of words from individual learners have the disadvantage of being less balanced than those that incorporate a smaller number of words from more learners, as individual usage may easily distort the overall results. For this reason, the contribution of individual learners is usually kept small in learner corpora (Nesselhauf, 2005: 43).*

Las principales ventajas son bien conocidas. En primer lugar, la recogida de datos en las que se basan el AE no suelen ser sistemáticas –frente a la recopilación de los datos de un corpus que sí lo son–: los detalles de los aprendientes y las circunstancias de la producción no se recogen o no lo suficientemente, y esta información es necesaria para realizar una adecuada interpretación de los datos, esto es, un buen análisis de la interlengua (Nesselhauf, 2005: 41-42). En segundo lugar, como ya señalaron Schachter y Celce-Murcia en 1977, la colección de textos se concibe como un depósito de errores que se desecha una vez que estos son extraídos, por lo que no se cuenta con el contexto para verificar resultados. Tampoco se atiende, como observa Nesselhauf (2005: 41-42), a la lengua que los aprendientes producen correctamente, hecho denunciado por Svartvik en 1973; los errores son contabilizados como absolutos y no suelen compararse con los aciertos en esa misma estructura o elemento. Asimismo, ni la sobreutilización ni la infrautilización pueden analizarse con este método. Con las siguientes palabras lo expone Leech (1998: xvii): el corpus de aprendientes

> *enables us to investigate the non-native speaking learner's language (in relation to the native speakers') not only from a negative point of view (what did the learner get wrong?) but from a positive one (what did the learner get right?). For the first time it also allows a systematic and detailed study of the learner´s linguistic behaviour from the point of view of "overuse" (what linguistic features does the learner use more than a native speaker?) and "underuse" (what features does the learner use less than a native speaker?).*

Asimismo, la consolidación de la lingüística de corpus, la evolución tecnológica y los métodos desarrollados en la investigación permitieron que los nuevos corpus informatizados pudieran ser desarrollados como herramientas de investigación para ser usadas por muchos especialistas en este campo, y no solo por un investigador individual como en el análisis del material usado en el periodo anterior; este es un aspecto clave del éxito de la investigación en ASL basada en corpus. Además, por medio del ordenador, se pueden realizar nuevos tipos de estudios, como obtener la frecuencia de coaparición –o estudios cuantitativos en general–, y descubrir patrones de uso lingüístico en un grupo concreto de aprendientes. Hay diferentes tipos de *software* de recuperación de los datos en función de los propósitos del investigador. Micro Concord, Word Smith Tools o AntConc son algunas de las herramientas de recuperación más usadas.

No obstante, pese a estos beneficios en el análisis por la informatización de los corpus, los investigadores no deberían limitar sus investigaciones a lo que el ordenador puede hacer, pues, como señala Granger (1998b: 16), "a computerized approach has linguistic limitations". Por un lado, las diferencias o las semejanzas superficiales entre los aspectos de la lengua nativa y no nativa siempre requieren investigación cualitativa (Meunier, 1998: 36). Dado que los hallazgos cuantitativos han de ser considerados cuidadosamente y comparados con análisis cualitativos, se plantea en Hasselgard y Johansson (2011: 55-56) la necesidad de utilizar corpus alternativos de referencia para contrastar los datos y buscar causas externas para algunos de los resultados. Además, 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), 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 y 2013):

> *The triangulation of corpus methods with other research methodologies will be an important further step in enhancing both the rigour of corpus linguistics and its incorporation into all kinds of research (…) To put it another way, the way ahead is methodological pluralism (McEnery y Hardie, 2012: 227).*

De este modo, hay que considerar los datos cuantitativos como punto de partida para un análisis posterior, por lo que filtrar los datos obtenidos por medios informáticos debería ser la esencia de la investigación de corpus de aprendientes.

En suma, se ha visto a lo largo de este recorrido por los diferentes modelos de análisis de la lengua del aprendiente que el AE no puede proporcionar una descripción completa de la lengua del aprendiente (recuérdese la reivindicación de Svartvik [1973: 8]), aunque sí intentó presentar el panorama de los errores de los aprendientes con claros propósitos pedagógicos; por otro lado, se ha puesto de manifiesto que el aspecto colectivo de la lengua del aprendiente ha recibido poca atención en la investigación actual sobre la ASL. Como se ha señalado, los corpus de aprendientes informatizados pueden suplir estas carencias, y sumar otras ventajas, aunque al mismo tiempo conviene conocer las limitaciones de esta metodología, a las que también nos hemos referido.

4. Bibliografía

ALONSO RAMOS, M. *et al*. (2010a): "Tagging collocations for learners". En S. Granger y M. Paquot (eds.), *eLexicography in the 21st Century: New Challenges, New Applications: Proceedings of ELEX2009, Cahiers du CENTAL 7*. Lovaina la Nueva: Presses universitaires de Louvain.

ALONSO RAMOS, . *et al*. (2010b). "Towards a motivated annotation schema of collocation errors in learner corpora". En N. Calzolari (ed.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (*LREC'10*). Valetta: Language Resources Evaluation.

ALTENBERG, B. (2011): "Preface". En F. Meunier *et al*., *A Taste for Corpora. In honour of Sylvianne Granger*. Amsterdam: John Benjamins, xiii-xv.

ALTENBERG, B. y S. GRANGER (2001): "The grammatical and lexical patterning of MAKE in native and non-native student writing". *Applied Linguistics,* 22 (2),

173-195.
BARALO, M. (2004 [1999]): *La adquisición del español como segunda lengua.* Madrid: ArcoLibros.
BOSQUE, I. (2004): "Combinatoria y significación. Algunas reflexiones", En I. Bosque (dir.), *Redes. Diccionario combinatorio del español contemporáneo*. Madrid: SM, LXXVII-CLXXIV.
BOSQUE, I., dir. (2004): *Redes. Diccionario combinatorio del español contemporáneo*. Madrid: SM.
BOSQUE, I., dir. (2006): *Práctico. Diccionario combinatorio del español contemporáneo*. Madrid: SM.
CORDER, Stephen Pit (1967): "The significance of learners' errors". *International Review of Applied Linguistics* 5: 161-170.
DAVIES, Mark y Dee Gardner (2010): *A Frequency Dictionary of Contemporary American English.* Londres: Routledge.
DE ALBA, Virginia (2009): "El análisis de errores en el campo de ELE. Algunas cuestiones metodológicas". *Revista Nebrija de Lingüística aplicada a la enseñanza de lenguas,* 5.
DULAY, H. *et al.* (1982): *Language two*. Oxford: Oxford University Press.
ELLIS, R. (1994): *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
ENKVIST, N. E. (1973): "Should we count errors or measure success?". En J. Svartvik, *Errata: Papers in error analysis*. Lund: Gleerup/Liber, 16-23.
FERNÁNDEZ, S. (1997): *Interlengua y análisis de errores*. Madrid: Edelsa.
FERNÁNDEZ GONZÁLEZ, J. (1995): *El análisis contrastivo: historia y crítica.* Valencia: Lynx, Universidad de Valencia.
FIRTH, J, R. (1957): *Papers in Linguistics 1934-1951*. Londres: Oxford University Press.
GRANGER, S. (1998a): "Introduction". En S. Granger (ed.), *Learner English on Computer*. Londres: Longman, XXI-XXII.
GRANGER, S. (1998b): "The computer learner corpus: a versatile new source of data for SLA research". En S. Granger (ed.), *Learner English on Computer*. Londres: Longman, 3-18.
GRANGER, S. *et al.,* eds. (2002): *Computer Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam: John Benjamins.
GRANGER, Sylviane (2004): "Computer learner corpus research: current status and future propects". En U. Connor y T. A. Upton (eds.), *Applied Corpus Lingusitics: A Multidimensional Perspective*. Amsterdam: Rodopi, 123-145.
GRANGER, S. *et al.,* eds. (2009): *International Corpus of Learner English*. Version 2. Lovaina la Nueva: Presses universitaires de Lovaina.
GILQUIN, G. y Stephan G. (2009): "Corpora and experimental methods: a state-of-the-

art review". *Corpus Linguistics and Linguistic Theory* 5 (1): 1-26.

GUO, X. (2006): *Verbs in the Written English of Chinese Learners: A Corpus-based Comparison between Non-native Speakers and Native Speakers*. Tesis doctoral, Universidad de Birmingham [en línea] <http://etheses.bham.ac.uk/871/> [consulta: 20-11-2014].

HASSELGARD, H. y S. JOHANSSON (2011): "Learner corpora and contrastive interlanguage analysis". En F. Meunier *et al*. (eds.), *A Taste for Corpora. In honour of Sylviane Granger*. Amsterdam: John Benjamins, 33-61.

HALLIDAY, Michael A.K. (1966): "Lexis as a linguistic level". En C. E. Bazell *et al.* (eds.), *In Memory of John Firth*. London: Longman, 148-162.

HALLIDAY, Michael A.K. (1977): "Estructura y función del lenguaje". En J. Lyons, *Nuevos horizontes de la lingüística*. Madrid: Alianza, 145-173.

HAMMARBERG, B. (1973): The insufficiency of error analysis. En J. Svartvik (ed.), *Errata: Papers in error analysis*. Lund: Gleerup/Liber, 29-36.

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

KRASHEN, Stepehen (1977): "Some issues relating to the monitor model". En H. D. Brown *et al.* (eds.)*, TESOL 77: teaching and learning English as a second language. Trends in research and practice*. Washington: TESOL, 144-158.

KUCERA, H. y W. N. FRANCIS (1967): *Computational Analysis of Present-day American English*. Providence RD: Brown University Press.

LADO, R. (1971 [1951]): *Linguistics across Cultures: Applied Linguistics for Language Teachers*. Ann Arbor MI: University of Michigan Press.

LARDIERE, D. (1998): "Dissociating syntax from morphology in a divergent L2 end-state grammar". *Second Language Research,* 14 (4), 359-375

LARSEN-FREEMAN, D. y M. H. LONG (1994): *Introducción al estudio de la adquisición de segundas lenguas*. Madrid: Gredos.

LEECH, Geoffrey (1998): "Preface". En S. Granger (ed.), *Learner English on Computer*. Londres: Longman, XIV-XX.

LEECH, G. (2001): "The role of frequency in ELT: new corpus evidence brings re-appraisal". *Foreign Language Teaching and Research,* 33 (5), 328-339.

LEECH, G. (2011): "Frequency, corpora and language learning". En F. Meunier *et al*. (eds.), *A Taste for Corpora*. *In honour of Sylvianne Granger*. Amsterdam: John Benjamins, 7-31.

LEWIS, M. (2000): *Teaching collocation: Further Developments in the Lexical Approach.* Hove: Language Teaching Publications.

LINNARUD, M. (1986): *Lexis in Composition: A Performance Analysis of Swedish Learners' Written English*. Lund: Gleerup/Liber.

LOZANO, C. (2009a): "CEDEL2: Corpus Escrito del Español L2". En C. M. Bretones Callejas *et al.* (eds.), *Applied Linguistics Now: Understanding Language and Mind.* Almería: Universidad de Almería, 197-212.

LOZANO, C. (2009b): "Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus". En Y. Leung *et al*., *Representational Deficits in Second Language Acquisition*. Amsterdam: John Benjamins, 127-166.

LOZANO, C. y A. MENDIKOETXEA (2013): "Learner corpora and SLA: the design and collection of CEDEL2". En A. Díaz-Negrillo *et al*. (eds.) *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, 65-100.

MAIRAL USÓN, R. (2010): *Teoría Lingüística: Métodos, Herramientas y Paradigmas*. Madrid: Fundación Ramón Areces/ UNED.

MCENERY, T. y A. HARDIE (2012): *Corpus Linguistics*. Cambridge: University Press.

MENDIKOETXEA, A. y C. LOZANO (2012): "On the need to combine corpus data and experimental data in L2 acquisition research". Conferencia presentada en el IV Congreso Internacional de Lingüística de Corpus. Universidad de Jaén.

MEUNIER, F. (1998): "Computer tools for the analysis of learner corpora". En S. Granger (ed.), *Learner English on Computer*. Londres: Longman, 19-37.

MUÑOZ LICERAS, J. (2009): "La interlengua del español en el siglo XXI". *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas,* 5.

NESSELHAUF, N. (2004): "Learner corpora and their potential for language teaching". En J. Sinclair (ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 125-152.

NESSELHAUF, Nadja (2005): *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

OROL GONZÁLEZ, A. y M. ALONSO RAMOS (2013): "A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish". *Procedia-Social and Behavioral Sciences,* 95: 563-570.

PASTOR CESTEROS, S. (2004): *Aprendizaje de segundas lenguas. Lingüística aplicada a la enseñanza de idiomas*. Alicante: Universidad de Alicante

PASTOR CESTEROS, S. (2005): "La enseñanza de segundas lenguas". En A. López y B. Gallardo (eds.), *Conocimiento y lenguaje*, Valencia: Universitat de València, 361-399.

PENADÉS, I., coord. (1999): *Lingüística contrastiva y análisis de errores*. Madrid: Edinumen.

PENADÉS, I. (2003): "Las clasificaciones de errores lingüísticos en el marco del análisis de errores". *LinRed,* 1: 1-29.

PRIETO GONZÁLEZ, Sabela *et al.* (2009): "Córpora y enseñanza de lenguas: se buscan colocaciones". En P. Cantos Gómez y A. Sánchez Pérez (eds.), *A survey on corpus-based research.* Murcia: AELINCO, 336-373.

SÁNCHEZ RUFAT, A. (2014): "Rasgos de la competencia léxica del verbo". Revista

Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas, 17 [en línea] http://www.nebrija.com/revista-linguistica/numero-17-2014 [consulta: 29-05-2015].

SÁNCHEZ RUFAT, A. (2015): *El verbo dar en el español escrito de aprendientes de L1 inglés: estudio comparativo entre hablantes no nativos y hablantes nativos basado en corpus.* Tesis Doctoral, Universidad de Extremadura.

SÁNCHEZ RUFAT, A. y F. JIMÉNEZ CALDERÓN (2013): "Combinatoria léxica y corpus como input". *Language Design*, 14: 61-81.

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

SCHACHTER, J. y M. CELCE-MURCIA (1977): "Some reservations concerning error analysis". *TESOL,* 11 (4): 441-451.

SCHUMANN, J. (1976): "Second language acquisition: The pidginization hypothesis". *Language Learning,* 26: 391-408.

SELINKER, L. (1972): "Interlanguage". *International Review of Applied Linguistics,* 10 (2): 209-231.

SINCLAIR, J. M. (1966): "Beginning the study of lexis". En C. E. Bazell *et al.* (eds.), *In Memory of John Firth*. Londres: Longman, 410-30.

SINCLAIR, John M. (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

SINCLAIR, J. M. (2005): "How to build a corpus". En M. Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice.* Oxford: Oxbow books, 79-83.

SINCLAIR, J. M. *et al*. eds. (2004): *English Collocation Studies*: *The OSTI Report*. Londres: Continuum.

STOCKWELL, R. *et al.* (1965): *The Grammatical Structures of English and Spanish*. Chicago: University of Chicago Press.

SVARTVIK, J., ed. (1973): *Errata: Papers in Error Analysis*. Lund: Gleerup/Liber.

TONO, Y. (2003): "Learner corpora: Design, development and applications". En D. Archer *et al*. (eds.), *Proceedings of the 2003 Corpus Linguistics Conference*. Lancaster University: UCREL Technical Paper 16, 800-809.

TRIBBLE, C. (2011): "Revisiting apprentice texts". En F. Meunier *et al*. (eds.), *A Taste for Corpora*. *In honour of Sylvianne Granger*. Amsterdam: John Benjamins, 85-108.

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

WANNER, L. *et al*. (2013): "Annotation of Collocations in a Learner Corpus for Building a Learning Environment". En S. Granger *et al*. (eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Lovaina: Presses universitaires de Louvain.