Análisis contrastivo de interlengua y corpus de aprendientes: precisiones metodológicas

ANNA SÁNCHEZ RUFAT
Profesora Sustituta Interina
Universidad de Córdoba
Facultad de Filosofía y Letras
Plaza del Cardenal Salazar, 3
14071 Córdoba
E-mail: asrufat@uco.es

ANÁLISIS CONTRASTIVO DE INTERLENGUA Y CORPUS DE APRENDIENTES: PRECISIONES METODOLÓGICAS

RESUMEN: Este trabajo pone de manifiesto la relevancia que han adquirido los corpus de aprendientes informatizados (CAI) en el análisis de la interlengua. Se describen algunos de los corpus más relevantes orientados al estudio de la adquisición de lenguas con la pretensión de esclarecer las cuestiones metodológicas que implica la investigación con este tipo de corpus. Para ello, se especifican, en primer lugar, las necesidades que los corpus de aprendientes han venido a cubrir, y que explican, al fin y al cabo, su aparición y su consolidación; en segundo lugar, se detallan las especificidades que poseen los estudios de interlengua basados en corpus, así como los factores que han de ser tenidos en cuenta para su emprendimiento, y se examina el proceso en relación con el campo específico del análisis de la interlengua del español; por último, se ponen de relieve ciertas limitaciones que deben considerarse al aplicar la metodología presentada. Se concluye que el estudio de CAI es ya un terreno consolidado de la Lingüística Aplicada que se ha convertido en un paradigma fructífero, si bien se halla en sus inicios en el caso de la interlengua del español. Quedan, asimismo, muchas cuestiones metodológicas por resolver para que los investigadores de Adquisición de Segundas Lenguas se beneficien de los estudios de CAI.

PALABRAS CLAVES: corpus de aprendientes; adquisición de segundas lenguas; interlengua del español.

SUMARIO: 1. Introducción: los corpus de aprendientes. 2. Corpus de aprendientes y Adquisición de segundas lenguas 3. Investigación de corpus de aprendientes. 4. Corpus de aprendientes y análisis de la interlengua del español. 5. Algunas cuestiones metodológicas de la investigación basada en corpus. 6. Conclusiones.

CONTRASTIVE INTERLANGUAGE ANALYSIS AND LEARNER CORPORA: METHODOLOGICAL ISSUES

ABSTRACT: This article highlights the importance that the computer learner corpora (CLC) have acquired in the analysis of interlanguage or learner language. Some of the most relevant leaner corpora are described in order to clarify the methodological issues involved in this type of corpus research, especially regarding research on interlanguage in Spanish. To do this, the paper first identifies the needs met by learner corpora, which explains the birth of this field of study and its consolidation. Secondly, it outlines the specificities of the interlanguage studies based on corpora and the factors to be taken into account for carrying out this kind of study are detailed. This process is also discussed in relation to the specific field of analysis of Spanish interlanguage; Finally, it highlights some limitations that should be considered when applying the methodology presented. We conclude that the study of CLC is already an established field of applied linguistics that has become a fruitful paradigm, although it is in its infancy in the case of the Spanish interlanguage. Furthermore, many methodological issues remain unsolved. As a result, researchers of Second Language Acquisition cannot fully benefit from studies of CLC.

KEY WORDS: Learner corpora; Second Language Acquisition; Spanish Interlanguage.

SUMMARY: 1. Introduction: learner corpora. 2. Learner corpora and Second Language Acquisition. 3. Learner corpus research. 4. Learner corpora and interlanguage analysis in Spanish. 5. Methodological issues in learner corpus research. 6. Conclusions.

ANALYSE CONTRASTIVE DE L'INTERLANGUE ET LE CORPUS D'APPRENANTS: PRECISIONS METHODOLOGIQUES

RÉSUMÉ: Cette étude met en évidence l'importance qu'ont acquise les corpus d'apprenants informatisés dans l'analyse de l'interlangue ou de la langue d'apprentissage. Certains corpus d'importance sont décris afin de clarifier les questions méthodologiques qu'implique la recherche à l'aide de ce type de corpus. Pour cela, en premier lieu, nous spécifions les besoins pour lesquels les corpus d'apprenants sont spécialement utiles et qui expliquent finalement leur apparition et leur consolidation ; en deuxième lieu, nous détaillons les caractéristiques des études de l'interlangue liées au corpus, ainsi comme les facteurs à prendre en compte pour son élaboration, et nous examinons en détail le processus lié au champ spécifique de l'analyse de l'interlangue en espagnol. En dernier lieu, nous soulignons certaines limitations à prendre en compte dans la méthodologie présentée. Nous concluons que l'étu3de des corpus d'apprenants informatisés est un domaine consolidé dans la linguistique appliquée qui s'est converti en un paradigme fructueux, bien qu'il se trouve toujours dans sa phase initiale dans le cas de l'interlange de l'espagnol. En plus, il reste de nombreuses questions méthodologiques à résoudre afin que les chercheurs d'acquisition des langues secondes puissent profiter pleinement des études des corpus d'apprenants informatisés.

MOTS CLÉS: corpus d'apprenants; acquisition des langues secondes; interlangue de l'espagnol.

SOMMAIRE: 1. Introduction : les corpus d'apprenants. 2. Corpus d'apprenants et acquisition des langues secondes. 3. Recherche du corpus d'apprenants. 4. Corpus d'apprenants et analyse de l'interlangue de l'espagnol. 5. Quelques questions méthodologiques dans la recherche liée au corpus. 6. Conclusions.

Fecha de Recepción 21/01/2015
Fecha de Revisión 26/10/2015
Fecha de Aceptación 01/11/2015
Fecha de Publicación 01/12/2015

DOI: http://dx.doi.org/10.25267/Pragmalinguistica.2015.i23.11

Análisis contrastivo de interlengua y corpus de aprendientes: precisiones metodológicas[1]

ANNA SÁNCHEZ RUFAT

1. INTRODUCCIÓN: LOS CORPUS DE APRENDIENTES

Los comienzos de la Lingüística de Corpus (LC) se remontan a los años sesenta, 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, dado que se concibe necesariamente en soporte electrónico: "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).

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, contamos con grandes corpus de referencia como el Corpus del Español, de Davies (2002) —que pasará de 20 millones a dos billones de palabras de muestras de lengua actuales en 2016—, el Corpus de Referencia del Español Actual (CREA) —con más 160 millones de palabras en su última versión (2008)— y el recién creado Corpus del Español del Siglo XXI (CORPES XXI) —actualmente con 200 millones de palabras de muestras de lengua recientes que permiten obtener mejores descripciones sobre el uso que se hace del español durante el siglo XXI (2014)—, estos dos últimos a cargo de la Real Academia Española.

Por otro lado, el uso de corpus en la investigación sobre la adquisición de la lengua materna (L1) tampoco es nuevo; desde los años 70 es una práctica común en los estudios de la lengua del niño. Dentro de este ámbito, el corpus más extenso es el 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.

Ahora bien, los corpus de aprendientes informatizados no aparecieron hasta comienzos de los 90, cuando la tecnología y el análisis de corpus de hablantes nativos (HN) se encontraban relativamente desarrollados. En relación lógica con la ya expuesta definición de corpus de Sinclair (2005: 16), se entiende por *corpus de aprendientes* "electronic collections of authentic FL/SL textual data according to explicit design criteria for a particular

[1] Este trabajo se enmarca en el proyecto de investigación titulado Modelos y Representaciones Metateóricas en la Historia de la Lingüística (FFI2012-35802), cuya IP es Carmen Galán.

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; dichos criterios, establecidos también por Sinclair (2005), son sintetizados por Lozano y Mendikoetxea (2013: 7): 1) selección del contenido de acuerdo con criterios externos, como la función comunicativa de los textos, y no la lengua utilizada; 2) representatividad del corpus del estado de lengua seleccionado; 3) contraste con un corpus de control diseñado a tal efecto; 4) criterio estructural basado en la sencillez; 5) almacenamiento por separado de las etiquetas y del texto en bruto; 6) los textos deben comprender actos de habla completos, independientemente de su tamaño; 7) el diseño y la composición del corpus deben estar documentados; 8) equilibro entre las muestras de lengua oral y escrita si el corpus pretende incluir ambas variedades; 9) el control del tema en un corpus solo debe imponerse por el uso de criterios externos; y, finalmente, 10) los textos deben ser homogéneos, por lo que se deben descartar los que no son representativos de la variedad seleccionada. Granger (2002: 4) se refiere también a los principales objetivos de este tipo de corpus: 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 segundas lenguas (L2).

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; y, por otro lado, porque las necesidades planteadas en el ámbito de la Adquisición de Segundas Lenguas (ASL) desembocan en la confluencia de ambas ramas, por lo que los análisis de la interlengua –o lengua del aprendiente– basados en extensos corpus disfrutan a partir de entonces de una buena acogida en este campo de la ASL. A estas necesidades nos referimos a continuación.

2. Corpus de aprendientes y ASL

A partir de los años 80, aproximadamente, el campo de estudio de la ASL demanda nuevas herramientas por varias razones. Por un lado, los estudios basados en el análisis del comportamiento de los HN no informan de la dificultad de las estructuras específicas que deben ser enseñadas en el aula de lengua extranjera (LE), ni de su proceso de aprendizaje; por ello, Granger (1998: 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 HNN 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 HNN. A este respecto, Leech (2001:

339) defiende que los análisis de corpus de aprendientes permiten identificar áreas de dificultad que no se derivan a partir de un análisis de un corpus nativo. Así pues, para el desarrollo de la enseñanza de una L2 es importante disponer de corpus nativos para conocer lo que estos verdaderamente utilizan y contar con corpus de HNN para averiguar sus dificultades.

Por otro lado, se hacen necesarios los corpus de aprendientes informatizados por las ventajas que ofrecen con respecto a los procedimientos adoptados normalmente en el Análisis de Errores (AE) tradicional (método de análisis de la lengua del aprendiente inmediatamente anterior a la investigación de Corpus de Aprendientes Informatizados [CAI]), 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 HNN deben producir–. Las principales ventajas de los corpus son bien conocidas. En primer lugar, 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; solo así se evita que el uso individual distorsione los resultados generales. En segundo lugar, la recogida de datos en la que se basa el AE no suele ser sistemática –frente a la recopilación de los datos de un corpus que sí lo es–: los detalles referentes a los HNN 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 tercer lugar, como ya señalaron Schachter y Celce-Murcia en 1977, la colección de textos en el AE 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 a la lengua que los HNN producen correctamente, hecho indicado 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 LC, 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 de las herramientas informáticas de recuperación de datos 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 HNN.

No obstante, pese a estos beneficios en el análisis de la interlengua por la informatización de los corpus, los hallazgos cuantitativos han de ser considerados cuidadosamente y comparados con análisis cualitativos, esto es, las diferencias o las semejanzas superficiales entre los aspectos de la lengua nativa y no nativa siempre requieren investigación interpretativa, pues las reflexiones teóricas deben surgir de los resultados estadísticos obtenidos de una investigación cuantitativa. Solo así es posible aprovechar las ventajas de la tecnología de la informática actual y la práctica y las teorías de la LC y de otras áreas de la ASL, como la psicolingüística y la enseñanza de LE.

3. Investigación de corpus de aprendientes

El interés por los corpus de aprendientes se desarrolla rápidamente en los años noventa y primera década del siglo XXI, inspirado por el trabajo iniciado por Granger y su equipo en el CECL (Centre for English Corpus Linguistics), el ICLE (International Corpus of Learner English), primer corpus de aprendientes reconocido internacionalmente[2]. Este proyecto ha estimulado la creación de corpus similares en otras interlenguas, ha generado una gran cantidad de estudios académicos y ha influido enormemente en el campo de las publicaciones en enseñanza del inglés (como las de Longman y Cambridge University Press –a las que nos referimos en el apartado 3– que, pese a usar sus propios corpus de HNN, reciben la influencia del ICLE directa o indirectamente) y en el de la ASL (McEnery y Hardie, 2012: 82).

Así, puede decirse que la aparición de la investigación de Corpus de Aprendientes Informatizados (CAI) se produce en 1998 con la publicación de *Learner English on Computer*, editado por Granger, una colección de artículos pioneros sobre la lengua del aprendiente, basada en el ICLE. El volumen coeditado por Granger (Granger *et al.*, 2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* es una continuación del anterior con más desarrollos en el campo. El ICLE es un corpus informatizado disponible comercialmente de 3,3 millones de palabras procedentes de ensayos argumentativos escritos por aprendientes de inglés con dieciséis L1 diferentes (3,7 millones de palabras en la versión del corpus

[2] Con anterioridad al trabajo realizado en el proyecto ICLE, se llevaron a cabo estudios basados en datos de corpus de aprendientes –entendiendo este concepto desde una perspectiva muy amplia– como el de Krashen *et al.* (1978), pero, como señalan McEnery y Hardie (2012: 82), el alcance y la naturaleza sistemática del trabajo de Granger se concibe como el primer compromiso a gran escala con la lengua del aprendiente por medio de un enfoque basado en corpus.

publicada en 2009). Pronto se completó con un corpus hablado (LINDSEI, Gilquin *et al.*, 2010), constituido por 800000 palabras de entrevistas a aprendientes de inglés de diferentes L1. Ambos corpus –anotados según la categoría de palabra y error– se corresponden únicamente con un diseño transversal, esto es, basado en distintos niveles de competencia. Para suplir esta limitación, se inició en 2008 un nuevo proyecto, el Longitudinal Database of Learner English (LONGDALE), para compilar un corpus longitudinal que integra HNN de nivel avanzado con diferentes L1 que son seguidos durante un periodo de tres años, por lo que este corpus puede revelar información sobre el propio proceso de aprendizaje.

Estos corpus de aprendientes desarrollados en el CECL permiten mostrar los errores y las desviaciones cuantitativas (sobreutilización e infrautilización) en los usos de los HNN a partir de una norma o estándar comparativo seleccionado[3]. Estos errores y desviaciones difieren en tipo y frecuencia dependiendo de la L1 del aprendiente, por lo que para poder determinar el papel que desempeña la transferencia de la L1 en la producción del aprendiente lo ideal es poder contar con un corpus multilingüe, como los anteriores, para desde una perspectiva contrastiva poder demostrar empíricamente las intuiciones en la teoría de ASL. Por ello, Granger propone en 1996 el Análisis Contrastivo de Interlengua (ACI) para analizar los corpus de aprendientes, y lo coloca en el centro del proyecto ICLE. La utilidad del ACI ha sido demostrada en una gran cantidad de estudios, como se refleja en Granger *et al.* (2002). A diferencia del estudio contrastivo, que implica normalmente la comparación de dos lenguas, el ACI involucra diferentes variedades de la misma lengua: "involves quantitative and qualitative comparisons between native language and learner language (L1 *vs.* [*versus*] L2) and between different varieties of interlanguage (L2 *vs.* L2)" (Granger, 2009: 18). No obstante, Hasselgard y Johansson (2011: 43) subrayan que la mayoría de los estudios de ACI realizados hasta ahora comparan la lengua nativa con una sola variedad no nativa, si acaso con dos; no en vano, Guo (2006: 22) mantiene que el ACI no solo se refiere a la comparación entre interlenguas diferentes, sino también a la comparación entre una interlengua particular y la lengua meta.

Los estudios de interlengua basados en corpus requieren, por lo tanto, un corpus nativo comparable, esto es, diseñado para permitir comparaciones. Este corpus de control sirve para medir las desviaciones de la lengua de los HNN con respecto a la nativa, aunque es posible realizar un análisis de interlengua basado en corpus no contrastivo; este se integra en los estudios de ASL, pero no en los de ACI; no obstante, esta no es una práctica estándar en la investigación de la L2 (Lozano y Mendikoetxea, 2013: 10). El

[3] Sobre las debilidades de esta norma comparativa descriptiva para determinar los errores y sobre la necesidad de establecer una norma prescriptiva, véase Sánchez Rufat y Jiménez Calderón (2013).

corpus de aprendientes suecos de Linnarud (1986), aunque no estaba informatizado, contaba ya con un corpus comparable de HN de inglés, por lo que puede ser considerado germen del ACI.

El corpus nativo de comparación con el que cuenta el ICLE es el Louvain Corpus of Native English Essays (LOCNESS), que contiene unas 324000 palabras procedentes de ensayos argumentativos escritos por estudiantes británicos y norteamericanos. Por su parte, el LINDSEI también dispone de un corpus nativo de control, el Louvain Corpus of Native English Conversation (LOCNEC). Todos estos corpus están suficientemente documentados, esto es, contienen información sobre el aprendiente (edad, sexo, L1, otras lenguas extranjeras conocidas, región, nivel de dominio, etcétera), datos necesarios en una investigación de interlengua basada en un corpus.

La influencia del ICLE se puede observar en numerosos proyectos en los que se utilizan corpus de interlengua del francés (FRIDA [Granger, 2003]), alemán (el corpus FALKO [Lüdeling *et al.*, 2005]), italiano (el corpus VALICO [Corino, 2008]) y noruego (el corpus ASK [Tenfjord *et al.*, 2006])[4]. Con el mismo diseño que el ICLE, el Written Corpus of Learner English o WriCLE (Rollinson y Mendikoetxea, 2010) es un corpus disponible gratuitamente de unas 750000 palabras de aprendientes de inglés con L1 español que ha sido creado en la Universidad Autónoma de Madrid, donde se está compilando un subcorpus de escritura no académica (WriCLEinf), en su mayoría blogs, para permitir la comparación entre los diferentes registros, así como el estudio de estructuras que no se encuentran normalmente en escritos académicos más formales.

Otros importantes corpus de aprendientes influidos también por el ICLE fueron compilados para satisfacer las necesidades de los creadores de materiales didácticos de consulta. Destaca el Longman Learners' Corpus (compilado por Longman Publishers), que pretende servir en la creación de diccionarios de enseñanza de lengua inglesa, como el *Longman Dictionary of Contemporary English* (1995), y para otros recursos de enseñanza y materiales didácticos que están diseñados para contrarrestar los errores hechos por estudiantes de determinadas L1. Este corpus cuenta con 10 millones de palabras escritas por HNN de 117 nacionalidades con diferentes niveles de dominio. El corpus está ahora disponible comercialmente, aunque ha originado muchas menos investigaciones que el ICLE.

El Cambridge Learners Corpus (Nicholls, 2003) es el mayor corpus de aprendientes de inglés –cuenta con 30 millones de palabras, y está en con-

[4] Véase http://www.uclouvain.be/en-cecl-lcworld.html (fecha de consulta: 20-06-14) para un extenso listado de diferentes corpus de aprendientes, integrado por varias interlenguas –en su mayoría inglés– y diferentes L1, con información sobre el tipo de datos que conforman el corpus, número de palabras y nivel de dominio del aprendiente.

tinuo desarrollo–, y también fue compilado para desarrollar material didáctico de inglés como lengua extranjera, en este caso, en la editorial Cambridge University Press; pero también está disponible comercialmente. Está creado a partir de más de 200000 exámenes escritos por aprendientes de inglés de 217 países diferentes. Contiene información sobre el aprendiente y las condiciones de redacción. Resulta de particular interés el hecho de que está anotado con la localización y tipos de errores cometidos por los HNN en su producción escrita. Las ventajas de disponer de un corpus anotado con los errores son, a menudo, destacadas:

> We can see which errors are typical of different learner levels or of particular language groups because all the scripts have information about the first language and English level of the writer. This means that when we produce a book designed for a particular level, e.g. Upper Intermediate, we can look at all scripts written by Upper Intermediate learners and very easily see exactly what mistakes they make. In this way we can make sure the book contains appropriate help for an Upper intermediate student[5].

Otros corpus de aprendientes, como The Hong Kong University of Science and Techonology Learner Corpus o The Chinese Learner English Corpus, por poner algún ejemplo más, han sido también utilizados para la investigación en estudios de la interlengua del inglés de sinohablantes (Chi *et al.*, 1994, o Guo, 2006, entre muchos otros); el primero consta de 25 millones de palabras escritas por aprendientes de inglés en bachillerato y universidad. Al segundo, formado por un millón de palabras, solo pueden acceder los usuarios del departamento de inglés de HKPU.

4. CORPUS DE APRENDIENTES Y ANÁLISIS DE LA INTERLENGUA DEL ESPAÑOL

Desde no hace mucho tiempo, contamos también con datos de la interlengua del español, lo cual no debe extrañar, dado el tremendo desarrollo que esta lengua ha experimentado en esta última década: "se ha producido un auge notable en lo que al estudio del español y a todo lo relacionado con las culturas hispánicas respecta" (Muñoz-Basols *et al.*, 2014: 1). Este hecho ha revertido en un creciente interés por la adquisición del español como segunda lengua, según se demuestra en trabajos como los de Pastor Cesteros (2001), Pérez Leroux y Muñoz Liceras (2002) o Montrul (2004), centrados en el análisis de aspectos morfosintácticos, o los más abarcadores de Lafford y Salaberry (2003) o Geeslin (2013); dicho interés, además, ha coincidido con la creación de varios corpus de aprendientes de español. No pretendemos aquí ser exhaustivos proporcionando un listado con todos los corpus de aprendientes de español, sino presentar los más destacados por su calidad en el diseño y, por lo tanto, por su potencial para servir como buena base de datos en las investigaciones de la interlengua del español. Además,

[5] Cita de la página *web* de Cambridge Universtiy Press recogida por McEnery y Hardie (2012: 83), aunque ya no se encuentra localizable.

como bien observan Lozano y Mendikoetxea (2013: 21), los investigadores están continuamente creando corpus para satisfacer sus necesidades, lo que dificulta un posible recuento de todos los corpus existentes en un momento dado.

Uno de los corpus del español más representativos por su rigor metodológico y tamaño es el CEDEL2 (Lozano, 2009a; Lozano y Mendikoetxea, 2013), que, inspirado en el ICLE, nace en el proyecto WOSLAC, de la Universidad Autónoma de Madrid y la Universidad de Granada, cuyo objetivo principal es el estudio del papel de las interfaces (léxico-sintaxis y sintaxis-discurso) como fuentes de error en el desarrollo de la interlengua del HNN. Es un corpus escrito de aprendientes anglófonos de español clasificados en un nivel de dominio lingüístico inicial, intermedio o avanzado, según los resultados obtenidos en un test de diagnóstico estandarizado. Este hecho es un dato importante para el análisis que, sin embargo, no es tenido en cuenta para la configuración del ICLE y de otros muchos corpus, donde se clasifican los participantes por medio de factores externos como el nivel de dominio que deberían tener de acuerdo con la edad y curso en el que se encuentran. Dichos factores no garantizan que los participantes de un corpus sean comparables en términos de dominio lingüístico, ni que los resultados de los análisis sean extrapolables a otros contextos de aprendizaje similares.

El CEDEL2 cuenta, asimismo, con un subcorpus nativo de control diseñado con las mismas características que el no nativo con el objetivo de servir como base de datos con la que comparar los usos de la interlengua. En marzo de 2011 (fecha del último recuento publicado) el corpus contaba con 750000 palabras, aunque se encuentra todavía en la fase de recopilación de datos; se pretende alcanzar el millón de palabras. Hasta ahora, los estudios realizados a partir del CEDEL2 han analizado algunos aspectos del español aislados, pertenecientes a diferentes niveles lingüísticos e interfaces. Así, Pino (2012) estudia el uso de *lo que*, *de que*, *algo que* y *dice que* en un grupo de aprendientes suecos de español comparado con el uso que hacen los HN del CEDEL2, y Lozano (2009b, 2009c) se ha centrado en el análisis de la adquisición de sujetos gramaticales. Algunas muestras del CEDEL2 han sido etiquetadas de acuerdo con la estructura sintáctica y colocaciones –por medio de UAM Corpus Tool (O' Donell, 2009)–. De hecho, se ha iniciado el estudio de las colocaciones en el CEDEL2 (Prieto *et al.*, 2009; Pérez Serrano, 2012; Orol González y Alonso Ramos, 2013) con vistas a diseñar aplicaciones en línea, asistentes informáticos que sirvan de herramienta de ayuda a la redacción en español, que detecten el error colocacional y aporten estrategias de corrección (Ferraro *et al.*, 2011; Vincze *et al.*, 2011; Wanner *et al.*, 2013a; Ferraro *et al.*, 2014), lo cual requiere la creación de una tipología del error colocacional que permita etiquetar el corpus de aprendientes (Alonso Ramos *et al.*, 2010a y 2010b; Wanner *et al.*, 2013b). En Sánchez Rufat (2015) se analizan detalladamente las colocaciones y otras combinaciones léxicas proyectadas por el verbo *dar* en el CEDEL2 a partir de varias técnicas

y procedimientos combinados: las relaciones de frecuencias, el test de significatividad y la tipología del error. Por otro lado, Escutia (2010 y 2012) analiza el uso de predicados inacusativos y del *se* por aprendientes de nivel avanzado.

El diseño del CEDEL2 atiende a los diez principios estándar recomendados por Sinclair (2005). El hecho de que un corpus esté bien diseñado permite poder contestar a cualquier pregunta de investigación sobre la adquisición del español; además, la redacción –que es el tipo de texto que configura el CEDEL2– es una actividad voluntaria para los HNN que carece de restricciones temáticas, ya que el participante elige de entre una gran variedad de temas. Esto hace que todas las estructuras y elementos léxicos estén suficientemente representados. Por eso, el CEDEL2 se considera una base de datos naturales.

El CEDEL2 es un buen complemento del Spanish Learner Language Oral Corpus (SPLLOC 1 y SPLLOC 2) (Mitchell *et al.*, 2008), que lanzaron poco antes la Universidad de Southampton, la Universidad de Newcastle y la Universidad de York con financiación del Economic and Social Research Council (ESRC) del Reino Unido. Este, a diferencia del CEDEL2, contiene datos orales, y no escritos, de anglohablantes clasificados en un nivel de dominio lingüístico en función de la edad y los años de estudio del idioma –y no a través de un test de diagnóstico estandarizado–, que aprenden español en contextos institucionales (secundaria y universidad). Este corpus también contiene datos de HN de español obtenidos a partir de las mismas entrevistas con objeto de poder comparar el español nativo y la interlengua española de los hablantes de inglés. Los datos se obtienen a partir de unas tareas de elicitación diseñadas para este proyecto, entre las que se incluye la narración, la descripción de imágenes, el debate sobre un tema determinado y la entrevista individual. Con los datos obtenidos de estas tareas se pretende contestar a las preguntas de investigación del proyecto, relacionadas con el uso de los pronombres clíticos y el orden de palabras en el caso de SPLLOC1, y con el tiempo y aspecto en el de SPLLOC2. Por consiguiente, a diferencia del CEDEL2, este corpus no contiene datos de producción naturales, sino *semi-naturales*, pues está diseñado para obtener determinados usos o estructuras que permiten al investigador dar respuesta a sus preguntas de investigación. Pese a que estas técnicas se desaconsejan según los criterios estándar de diseño de corpus, el SPLLOC supone un hito para la investigación de corpus de aprendientes de español, por ser el primero que está compilado de acuerdo con ciertos criterios estándar de diseño, lo que lo convierte en un corpus muy útil para la investigación en ASL.

Existen otros corpus de aprendientes de español, la mayoría de orientación pedagógica que a menudo se emplean para el análisis de errores. Destacan el Corpus de Aprendices Taiwaneses de Español (CATE; Lu, 2010), todavía en desarrollo y que ya cuenta con casi medio millón de palabras, o

The Japanese Learner Corpus of Spanish (Kamakura), con 83400 palabras escritas por aprendientes japoneses; de reciente creación es el Corpus Español de Aprendientes Italianos (SCIL; Bailini, 2013) y en el corpus The Anglian Polytechnic University Learner Spanish Corpus, de la Universidad Anglia Ruskin, en el Reino Unido (Ife, 2004)[6], los participantes son aprendientes de varias L1. Por otro lado, el corpus Aprender a Escribir en Lovaina, de la Universidad Católica de Lovaina, es un corpus escrito de 1 millón de palabras de aprendientes de L1 alemán de todos los niveles de dominio; su acceso en línea está restringido. En España contamos también con el Corpus para el Análisis de Errores de Aprendices de E/LE (CORANE) (Cestero Mancera *et al.*, 2001), creado con fines pedagógicos, que desde 2009 se distribuye en CD (Cestero Mancera y Penadés, 2009). Reúne más de mil escritos, etiquetados, de informantes de diferentes niveles de dominio y muy variadas L1, al igual que el recientemente compilado corpus escrito CAES (Instituto Cervantes, 2014; véase Parodi, 2015) y el corpus oral CORELE (Campillos Llanos, 2014); este último contribuye al análisis del error ayudado por las técnicas informatizadas (del inglés *Computer-aided Error Analysis*, Granger [2008]). Este enfoque emplea las técnicas de la LC y hunde sus raíces en el marco del AE propuesto por Corder (1971 [identificación, descripción y explicación del error]), pero requiere dar un paso más: la etiquetación de los errores de acuerdo con una descripción y clasificación del error previamente establecida; este corpus (CORELE) también cuenta con un corpus de control nativo, por lo que permite la comparación entre el español de los HNN y el de los HN[7].

De todo esto se deduce que ya existe material suficiente para poder investigar la lengua de los aprendientes de español, lo que permite enriquecer el panorama de los estudios de esta interlengua. Una vez que se diseñan corpus que reúnen las condiciones de fiabilidad y de estandarización de los datos, los cimientos para estrechar la relación entre la LC y la investigación de ASL ya están establecidos; si sobre ambos pilares se construye la investigación de CAI, este es el momento, por lo tanto, de realizar un sólido avance en esta rama de estudio de la lengua del aprendiente de español.

5. Algunas cuestiones metodológicas de la investigación basada en corpus

La investigación en la LC puede clasificarse en función de unos criterios que distinguen los tipos de corpus con los que puede trabajarse. A continuación, se presentan brevemente algunos de los criterios que consideramos más relevantes.

[6] Para más información sobre los corpus señalados en esta sección, véase n. 3.
[7] Para más información sobre diferentes enfoques basados en corpus para investigar la interlengua del español véase Campillos Llanos (2014) y Mendikoetxea (2013).

1) Uso de corpus anotado *versus* no anotado. Cuando el análisis lingüístico está codificado en los propios datos del corpus, se trata de un corpus anotado. En lugar de editar el texto directamente, es posible almacenar las anotaciones separadas del texto –para que el texto no se sobrecargue de etiquetas (Sinclair, 2004: 191)–, de tal manera que son los programas informáticos los que combinan e integran el texto y las anotaciones en función de lo que requiera el analista. Un corpus se puede anotar de modo que cada palabra tenga, por poner algún ejemplo, una etiqueta con la categoría gramatical a la que pertenece; se puede anotar el número y la persona de los verbos o pronombres, si el sustantivo tiene un referente animado o inanimado, o los errores. Al respecto de la etiquetación de estos últimos, son muchos los autores que defienden las cualidades de un corpus anotado; una vez que se diseña una taxonomía del error y se insertan las etiquetas en los archivos de texto, el corpus puede ser consultado de manera automática, y pueden obtenerse listas de tipos de errores específicos o las frecuencias de cada tipo de error.

Para automatizar la etiquetación del error se utiliza un *software* especial, como el Error Editor, usado en el CECL; el Knowator, aplicado al CEDEL2 en los errores colocacionales; o el Exchanger XML Lite 3.2. (eXtensible Markup Language). En Rayson y Baron (2011) se presenta una nueva aplicación, el llamado *Variant Detector*, para detectar los errores ortográficos en corpus escritos de HNN. De este modo, la explotación de los datos puede ser rápida si se cuenta con herramientas de búsqueda adecuadas que mejoren la precisión de las búsquedas. Para Leech (1997: 4-6), la anotación añade valor al corpus original porque permite la extracción directa de la información requerida por el analista y porque permite la reutilización del corpus por otros usuarios sin que estos tengan que reanotar el mismo material lingüístico.

Estas ventajas no implican que el proceso de etiquetación de errores esté exento de limitaciones: ni existe un estándar uniforme sobre lo que es correcto e incorrecto –los nativos tampoco parecen estar siempre de acuerdo en lo que es un error y lo que es aceptable en su lengua– ni, por tanto, son incuestionables los hallazgos obtenidos a partir de un corpus de HNN etiquetado con los errores[8]. Al mismo tiempo, no conviene perder de vista que un corpus etiquetado con los errores no permite descubrir otros aspectos de la interlengua que también contribuyen a caracterizar la lengua del aprendiente como lengua no nativa, como los sobreusos e infrausos –por lo que adolece de las mismas debilidades del AE, que se centraba en los usos incorrectos y no consideraba los usos correctos ni las estrategias de evitación.

En suma, la etiquetación del error, aunque puede resultar útil, no es una solución perfecta para el problema de lo no nativo en el uso de la L2. Es un tipo de información, entre toda la que puede proporcionar el corpus de HNN,

[8] Acerca de la selección de un estándar estable, véase Sánchez Rufat y Jiménez Calderón (2013).

que puede servir de ayuda al investigador de la interlengua, al estudiante y al profesor.

Como se ha adelantado, en lugar de editar el texto directamente, otra opción es contar con el texto libre de anotaciones externas –las cuales aparecen aisladas fuera del texto–, y esta posibilidad enfrenta a los especialistas. Los defensores del texto no anotado consideran que las etiquetas pueden imponer un análisis a los usuarios de los datos, o hasta pueden ser imprecisas o inconsistentes. Separar el texto de las etiquetas está en consonancia con la propuesta de Hunston (2002: 94) de no mezclar el texto original y la anotación, dado que debe haber un movimiento constante entre el uso de técnicas avanzadas de búsqueda en un corpus anotado y la visión de los datos lingüísticos originales, sin procesar.

2) Uso de un corpus longitudinal *versus* transversal. Los corpus diacrónicos, aplicados a la investigación de corpus de aprendientes, se conocen como *corpus longitudinales* (Granger, 2002: 11), usados para rastrear el desarrollo lingüístico de los HNN durante un periodo de tiempo. Estos son muy útiles para describir el progreso y la evolución de la interlengua. Debido a la dificultad a la hora de compilar un corpus de este tipo –que requiere un seguimiento de un grupo de HNN a lo largo de unos años–, existen pocos de esta naturaleza, como el LONGDALE, de la Universidad de Lovaina, o el LANGSNAP (Tracey-Ventura *et al.*, 2013) –este último para la interlengua del español–, que es una fuente reciente de datos longitudinales de entrevistas a aprendientes que estudiaron español en el extranjero. Los corpus longitudinales, al poder aportar información valiosísima acerca de la adquisición de la L2, son de gran interés para los investigadores, que intentan cubrir la carencia de corpus longitudinales por medio de la creación de corpus de aprendientes de diferentes edades o niveles (de inicial a avanzado), de manera que la estructura del corpus se asemeje a la de un corpus longitudinal. Este tipo de corpus se conoce como *quasi-longitudinal* (Granger, 2002: 11), y el CEDEL2 es un buen ejemplo al contener muestras de lengua de HNN de español de nivel inicial, intermedio y avanzado. No obstante, la mayoría de estudios realizados hasta ahora están basados en corpus de aprendientes sincrónicos, también llamados *transversales*; estos corpus permiten estudiar tendencias en un grupo de HNN, aunque son los longitudinales los que guardan una relación más estrecha con la investigación en ASL, ya que la principal ocupación de esta disciplina hasta la fecha ha sido la naturaleza del proceso de adquisición de la segunda lengua y los factores que afectan a la lengua de los aprendientes (Larsen-Freeman y Long, 1991).

3) Análisis de corpus cuantitativo y cualitativo. Los corpus son una fuente de datos cuantitativos para los lingüistas, por lo que a menudo sintetizan sus hallazgos cuantitativos a través de estadísticas. Es posible usar un corpus y no recurrir a análisis estadísticos, como cuando queremos es-

tablecer si un fenómeno particular existe en la lengua. Nos basta con recuperar una sola aparición de uso para probar que existe, aunque la no aparición en un corpus no implica la no existencia. En cambio, si lo que nos interesa es determinar si el fenómeno en cuestión es frecuente en la lengua, hay que emplear estadísticas. La frecuencia en los datos se produce con tanta regularidad en los análisis de corpus que es raro dar con un estudio en LC en el que no se lleve a cabo algún análisis estadístico (McEnery y Hardie, 2012: 49), aunque este sea relativamente básico y descriptivo. No obstante, cualquier estudio sobre la lengua del aprendiente resultaría superficial si el único análisis realizado consistiera en la recuperación de datos de manera mecanizada. Como ya se ha señalado, las diferencias o semejanzas superficiales entre la lengua nativa y no nativa siempre requieren una investigación posterior, por lo que los datos obtenidos de las búsquedas informatizadas suponen el punto de partida para el análisis ulterior, que ha de ser cualitativo. Esto no siempre sucede: como señala Muñoz Liceras (2009), en muchos casos se ha querido primar una investigación cuantitativa en el sentido más literal y menos útil del término.

Un análisis de corpus proporciona información sobre la interlengua y el proceso de aprendizaje; en relación con ello, señalamos otras cuestiones metodológicas que debemos tener en cuenta al realizar un estudio de CAI. Las inferencias sobre el proceso de aprendizaje fundamentadas únicamente en datos producidos de manera natural, como los que constituyen los corpus, deben ser interpretados como posibilidades, no certezas, pues los investigadores dependemos de los datos de producción, por lo que estamos restringidos a una cantidad de datos limitada, en la que solo se atiende a aquellos datos que pueden ser contados y se deja fuera los que no lo son. Además, los corpus son fuentes de datos que constituyen muestras de actuación, no de competencia, por lo que solo permiten deducir cómo es la competencia de forma indirecta; por ello, la ASL siempre requerirá otras fuentes de datos, como los experimentales, los metalingüísticos y los de la introspección (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:

> 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).

En algunos trabajos recientes sobre la interlengua del español esta triangulación metodológica ya se está aplicando; en estos estudios se combina el análisis de corpus y el análisis experimental psicolingüístico, como ocurre en Domínguez *et al.* (2013). Es preciso que el resto de los estudios sobre la interlengua del español terminen de girar en esa dirección. Por último, aunque los estudios de corpus estén centrados en las aplicaciones

pedagógicas de los resultados, debieran relacionarse con los debates, hipótesis y teorías actuales sobre la ASL y sus implicaciones en el desarrollo de la lengua del aprendiente; en la mayoría esto no sucede (Lozano y Mendikoetxea 2013).

6. Conclusiones

La investigación de la adquisición del español L2 basada en corpus está aún en una fase embrionaria; está prácticamente todo por hacer. Como se ha señalado, existen diferentes enfoques en los estudios de CAI; la mayoría comparan la interlengua de un grupo concreto de aprendientes con un corpus de control nativo de similares características, otros cuentan con un número de variedades de la misma interlengua representadas en el corpus o con la posibilidad de disponer de textos comparables en dos interlenguas diferentes. Los dos primeros tipos de corpus constituyen la base de datos de los estudios contrastivos de interlengua; el ICLE es uno de los corpus más representativos, ya que ofrece comparaciones entre dieciséis interlenguas del inglés. Este enfoque permite describir la interlengua en un nivel de dominio determinado atendiendo a los factores interlingüísticos (influencia de la L1) y a los intralingüísticos del aprendizaje, lo que posibilita un desarrollo enorme de los estudios sobre el proceso de adquisición de la L2 (en este caso, del inglés). La mayoría de los corpus de español no cuenta con un subcorpus comparable de esta misma interlengua producida por hablantes de diferentes L1, por lo que no se pueden realizar análisis de esta naturaleza. En cambio, el CEDEL2 sí que cuenta con un corpus comparable de una interlengua diferente, como es el caso del WriCLE. La comparación entre el CEDEL2 y el WriCLE permite determinar si las particularidades o déficits en una interlengua concreta (del español o del inglés) resultan de la influencia de la L1, si es consecuencia del *input* o de patrones de desarrollo universales (Lozano y Mendikoetxea, 2013: 6). El WriCle se encuentra todavía en fase de compilación de datos, por lo que queda pendiente su utilización en futuras investigaciones.

En suma, el estudio de CAI es ahora un campo establecido de la Lingüística Aplicada, un campo que está en continua evolución. Aunque el ACI se ha convertido en un paradigma fructífero, en el caso de la interlengua del español se está iniciando ahora su desarrollo. No obstante, mientras no se logre combinar corpus de HNN de español bien diseñados, buenas herramientas de búsqueda y anotaciones en el corpus, no será posible explorar todas las preguntas de investigación imaginables. Asimismo, tampoco se producirán importantes avances en el análisis de la interlengua del español si no buscamos una solución a las limitaciones ya referidas en torno a la metodología de análisis de corpus; solo así los investigadores de ASL se beneficiarán plenamente de los estudios de CAI.

REFERENCIAS

ALONSO RAMOS, M. (2010a): "Tagging collocations for learners", Granger, S. y Paquot, M. (eds.): *eLexicography in the 21st Century: New Challenges, New Applications: Proceedings of ELEX2009, Cahiers du CENTAL 7*, Lovaina la Nueva: Presses Universitaires de Louvain, pp. 375-380.

ALONSO RAMOS, M. (2010b): "Towards a motivated annotation schema of collocation errors in learner corpora", N. Calzolari (ed.): *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (*LREC'10*), Valetta, Malta: Language Resources Evaluation, pp. 3209-3214.

CAMPILLOS LLANOS, L. (2014): "A Spanish oral learner corpus for computer-aided error analysis", Corpora, 9 (2), pp. 207-238.

CESTERO MANCERA, A. M. *et al.* (2001): "Corpus para el análisis de errores de aprendinces de E/LE (CORANE)", *Actas del XII Congreso Internacional de ASELE: tecnologías de la información y de las comunicaciones en la enseñanza de la E/LE*, Publicación electrónica: http://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/12/12_0527.pdf.

CESTERO MANCERA, A. M. y PENADÉS, I. (2009): *Corpus de textos escritos para el análisis de errores de aprendices de E/LE (CORANE). CD-ROM*, Alcalá de Henares: Universidad de Alcalá.

CHI, M. *et al.* (1994): "Collocational problems amongst ESL learners: A corpus-based study", Flowerdew, L. y Tong A. K. (eds.): *Entering Text*, Hong Kong: University of Science and Technology, pp. 157–165.

CORDER, P. (1971): "Idiosyncratic dialects and error analysis", *International Review of Applied Linguistics,* 9, pp. 158-171.

CORINO, E. (2008): "VALICO: An Online Corpus of Learning Varieties of the Italian Language", Lyding, V. (ed.): *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics,* Publicación electronica: http://www.eurac.edu/en/research/autonomies/commul/conferences/Documents/LULCL II 2008_web_publication.pdf, pp. 117-134.

DOMÍNGUEZ, L. *et al.* (2013): "The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology", *Bilingualism: Language and Cognition,* 16 (3), pp. 558-577.

ESCUTIA, M. (2010): "El uso de *se* con verbos inacusativos por estudiantes avanzados de español como lengua extranjera", *Resla*, 23, pp. 129-151.

ESCUTIA, M. (2012): "Expletives and Unaccusative Predicates in L2", *Higher Education of Social Science,* 2 (3), pp. 1-14.

FERRARO, G. *et al.* (2011): "Collocations: A Challenge in Computer Assisted Language Learning", Boguslavsky, I. y Wanner, L. (eds.): *Proceedings of the Fifth International Conference on Meaning-Text Theory,* Publicación electrónica: http://olst.ling.umontreal.ca/pdf/proceedingsMTT2011.pdf, pp. 69-79.

FERRARO, G. *et al.* (2014) "Towards advanced collocation error correction in Spanish learner corpora", *Language Resources and Evaluation,* 48 (1), pp. 45-64.

GEESLIN, K. (ed.) (2013): *The Hand-*

book of Spanish Second Language Acquisition, Malden, Massachusetts: Blackwell/John Wiley.

GILQUIN, G. *et al.* (2010): *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*, Lovaina la Nueva: Presses Universitaires de Louvain.

GRANGER, S. (ed.) (1998): *Learner English on Computer*, Londres: Longman.

GRANGER, S. (2002): "A bird's-eye view of learner corpus research", Granger, S. *et al.* (eds.): *Computer Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam: Benjamins, pp. 3-33.

GRANGER, S. (2003): "Error-tagged learner corpora and CALL: a promising synergie", *CALICO*, 20 (3), pp. 465-480.

GRANGER (2008): "Learner corpora", Lüdeling, A. y M. Kytö (eds.): *Corpus Linguistics: An International Handbook*, (Volumen 1), Berlin: Mouton de Gruyter, pp. 259-75.

GRANGER, S. (2009): "The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation", Aijmer, K. (ed.): *Corpora and Language Teaching*, Amsterdam: Benjamins, pp. 13-32.

GRANGER, S. *et al.* (eds.) (2002): *Computer Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam: Benjamins.

GUO, Xiaotian (2006): *Verbs in the Written English of Chinese Learners: A Corpus-based Comparison between Non-native Speakers and Native Speakers,* tesis doctoral de la Universidad de Birmingham, *The linguistics journal*, Publicación electrónica: http://www.linguistics-journal.com/thesis_Guo.pdf.

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

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

INSTITUTO CERVANTES (2014): Corpus de Aprendices de Español (CAES), Publicación electrónica: http://galvan.usc.es/caes.

KRASHEN, S. *et al.* (1978): "Two studies in language acquisition and language learning", *ITL Review of applied linguistics*, 39/40, pp. 73-92.

LAFFORD, B. y SALABERRY, R. (2003): *Spanish Second Language Acquisition: State of the Science,* Washington: Georgetown University Press.

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

LEECH, G. (1997): "Introducing corpus annotation", Garside, R. *et al.* (eds.): *Corpus annotation: Linguistic information from computer text corpora*, Londres: Longman, pp. 1-18.

LEECH, G. (1998): "Preface", S. Granger, S. (ed.): *Learner English on Computer*, London: Longman, pp. XIV-XX.

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

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", Bretones Callejas, C. M. *et al.* (eds.): *Applied Linguistics Now: Under-*

standing Language and Mind, Almería: Universidad de Almería, pp. 197-212.

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

LOZANO, C. (2009c): "Pronominal deficits at the interface: New data from the CEDEL2 corpus, C. Bretones, C. *et al.* (eds.): *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada Hoy: Comprendiendo el lenguaje y la Mente*, Almería: Universidad de Almería, pp. 213-227.

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

LÜDELING, A. *et al.* (2005): "Multilevel Error Anotation in Learner Corpora", *The Corpus Linguistics Conference Series 1 (1) Corpus Linguistics,* Publicación electrónica: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/pdf/FALKO-CL2005.pdf.

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

MENDIKOETXEA, A. (2013): "Corpus-based research in second language Spanish", K. L. Geeslin: *The Handbook of Spanish Second Language Acquisition*, Oxford: Wiley-Blackwell, pp. 11-29.

MITCHELL, R. *et al.* (2008): "SPLLOC: A new corpus for Spanish second language acquisition research", Roberts, L. *et al.* (eds.): *EUROSLA Yearbook 8*, Amsterdam: Benjamins, pp. 287-304.

MONTRUL, C. (2004): *The Acquisition of Spanish: Morphosyntactic Development in Monolingual and Bilingual L1 Acquisition and Adult L2 Acquisition*, Amsterdam: Benjamins.

MUÑOZ-BASOLS, J. *et al.* (2014): "Hacia una inaternacionalización del discurso sobre la enseñanza del español como lengua extranjera", *Journal of Spanish Language Teaching*, 1 (1), pp. 1-14.

MUÑOZ LICERAS, J. (2009): "La interlengua del español en el siglo XXI", *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, 5, Publicación electrónica: http://www.nebrija.com/revistaliguistica/files/articulosPDF/articulo_5316fc5066e5c.pdf.

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

NICHOLLS, D. (2003): "The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT", Publicación electronica: <http://ucrel.lancs.ac.uk/publications/CL2003/papers/nicholls.pdf> .

O'DONELL, M. (2009): "The UAMCorpusTool: software for corpus annotation and exploration", Bretones C. *et al.*: *Applied Linguistics Now: Understanding Language and Mind/La Linguïstica Aplicada actual: Comprendiendo el Lenguaje y la Mente*, Almería: Universidad de Almería, pp.197–212.

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

PARODI, G. (2015): Reseña del Corpus de aprendices de español (CAES), Rojo G. y Palacios, I (dirs.): *Journal of Spanish Language Teaching* 2 (2).

PASTOR CESTEROS, S. (2001): "La concordancia en la interlengua de los aprendices de español como lengua extranjera", Pastor Cesteros, S. y Salazar, V. (eds.): *Tendencias y Líneas de Investigación en adquisición de segundas lenguas. Anexo I*, Alicante: Universidad de Alicante, pp. 5-60.

PÉREZ LEROUX, A. y MUÑOZ LICERAS, J. (2002): *The Acquisition of Spanish Morphosyntax: The L1/L2 Connection*, Dordrecht: Kluwer.

PÉREZ SERRANO, M. (2012): *Buscando colocaciones: análisis de errores colocacionales en un corpus de aprendientes de español.* Trabajo de doctorado: Universidad de Salamanca.

PINO, A. (2012): *El uso de combinaciones de palabras con que en un corpus de aprendices suecos de español como lengua extranjera,* tesis doctoral de la Universidad de Gotemburgo.

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

RAYSON, P. y BARON, A. (2011): "Automatic error tagging of spelling mistakes in learner corpora", F. Meunier *et al.* (eds.): *A Taste for Corpora. In honour of Sylvianne Granger*, Amsterdam: Benjamins, pp. 109-126.

RINGBOM, H. (1998): "High frequency verbs in the ICLE corpus", Renouf, A. (ed.): *Explorations in Corpus Linguistics*, Amsterdam: Rodopi, pp. 191-200.

ROLLINSON, P. y Mendikoetxea, A. (2010): "Learner corpora and second language acquisition: Introducing WriCLE", Bueno Alonso, J. L. *et al.* (eds.): *Analizar datos > Describir variación / Analysing Data > Describing Variation*, Vigo: Universidad de Vigo, pp. 1-12.

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 de la Universidad de Extremadura.

SÁNCHEZ RUFAT, A. y JIMÉNEZ CALDERÓN, F. (2013): Apreciaciones sobre la cuestión de la norma en el análisis de la interlingua, *Normas: Revista de Estudios Lingüísticos Hispánicos,* 3, pp. 183–204.

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

SINCLAIR, J. M. (ed.) (2004): *How to use corpora in language teaching*, Amsterdam: Benjamins.

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

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

TENFJORD, K. *et al.* (2006): "The ASK corpus: A language learner corpus of Norwegian as a second language", Proceedings of the 5th International Conference on Language Resources and Evaluation, pp. 1821-1824.

TRACEY-VENTURA, N. *et al.* (2013): "A longitudinal learner corpus investigation of vocabulary learning before, during, and after residence abroad*", Learner Corpus Research Conference*, Bergen, pp. 27-29.

VINCZE, O. *et al.* (2011): "Exploiting a

learner corpus for the development of a CALL environment for learning Spanish collocations", Kosem, I. y Kosem, K. (eds.): *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex,* Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 280-285.

WANNER, L. *et al.* (2013a): "Writing assistants and automatic lexical error correction: word combinatorics", *Proceedings of eLex 2013*, pp. 472-487.

WANNER, L. *et al.* (2013b): "Annotation of Collocations in a Learner Corpus for Building a Learning Environment", Granger, S. *et al.* (eds.): *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*, Lovaina: Presses universitaires de Louvain, Publicación en línea: http://lucas.dc.fi.udc.es/app/webroot/files/file/LCR2011_proceedings_wanner_leo_1_.pdf