linred

Lingüística en la red. Koelpucsjseqilgre 05 / 01 / 2016
eiflpqszcmIpwAnna.Sánchez.Rufataimlbiapmwqapdi
prugkñaqUniversidad.de.Extremadurawñubei fñoiyl
dlyrtlmvlinwytisdñfbxcieghñbvotlaqwiopyjhwxbzci

.artículo

Las funciones diagnóstica y evaluativa del análisis contrastivo de la interlengua del español basado en corpus

The diagnostic and evaluative functions in the contrastive interlanguage analysis based on corpus

Resumen

En la disciplina de la Adquisición de Segundas Lenguas, uno de los métodos que más contribuye a la investigación de la lengua del aprendiente es el Análisis Contrastivo de la Interlengua basado en corpus. Su aportación se basa fundamentalmente en sus dos funciones principales: la evaluativa y la diagnóstica. La importancia de estas radica en que revelan el aspecto colectivo de la naturaleza de la interlengua, el cual debería constituir una parte fundamental de la investigación de la Adquisición de Segundas Lenguas. No obstante, el aspecto colectivo de la interlengua del español no ha sido suficientemente estudiado. En este trabajo se examinan las funciones mencionadas y se proporcionan técnicas y procedimientos combinados que, aplicados a los análisis contrastivos de la interlengua basados en corpus, mejoran las descripciones de la lengua del aprendiente; en concreto, nos centramos en el enfoque de análisis que consiste en contrastar la producción lingüística de los hablantes no nativos con la de los usuarios expertos y/o nativos en un contexto de producción semejante y, por tanto, comparable.

Palabras Clave

Análisis Contrastivo de la Interlengua, investigación de corpus de aprendientes, interlengua del español, adquisición de ELE, corpus y ELE, función diagnóstica, función evaluativa

Abstract

This article introduces Contrastive Interlanguage Analysis (CIA), a corpus-based method for studying interlanguage in Second Language Acquisition (SLA). CIA's main contribution to SLA research can be found in its two main functions, namely, evaluative and diagnostic. CIA's importance lies in revealing the collective aspect of Spanish interlanguage. This has not yet been sufficiently studied, although it should be a facet of SLA research, as collective items are generalizable to all non-native speakers with the same background. This article discusses both functions and provides several techniques and combined procedures that can be applied to contrastive interlanguage studies based on an approach which consists in contrasting the production of no native and expert and/or native speakers of a language in a similar and, therefore, comparable situation.

Key words

Contrastive Interlanguage Analysis, learner corpus research, function of the corpus, Spanish interlanguage, acquisition of ELE, corpus and Spanish as a foreign language teaching, diagnostic function, evaluative function

Fecha de recepción: 09/09/2015 - Fecha de aceptación: 06/11/2015 – Fecha de publicación: 05/01/2016

1. Introducción: Investigación de Corpus de Aprendientes y Análisis Contrastivo de la Interlengua

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

Uno de estos procedimientos desarrollados para la obtención de datos naturales constituye la base del área de investigación lingüística conocida como Investigación de Corpus de Aprendientes o ICA (del inglés *Learner Corpus Research* o LCR[2]), que surge a finales de los años 80 como resultado de la interrelación de dos disciplinas hasta entonces desvinculadas entre sí: la Lingüística de Corpus (LC) y la ASL (Granger 2002: 4). El desarrollo de los corpus de aprendientes informatizados y el de la Investigación de Corpus de Aprendientes ha supuesto un gran avance en los estudios de ASL. Mientras que los trabajos anteriores –basados en otros procedimientos de obtención de datos– eran limitados en cuanto al número de sujetos estudiados para poder controlar las variables que afectan a la producción del aprendiente, los nuevos corpus informatizados permiten trabajar con más y mejor calidad de datos de lengua natural. Así pues, poder analizar amplios corpus de aprendientes que estén bien diseñados, de acuerdo con los criterios establecidos (Sinclair 2005), proporciona una base empírica sólida para la descripción de la interlengua (IL). Al mismo tiempo, estas descripciones mejoradas de la lengua del aprendiente pueden ser utilizadas en la enseñanza de las L2 (Granger 2002: 4; Pastor Cesteros 2004: 98).

Sylviane Granger ha sido, sin duda, uno de los principales motores de la Investigación de Corpus de Aprendientes (véase Granger 1996). En los años 90 inició, en la Universidad Católica de Lovaina, dos proyectos pioneros que han marcado, por su alcance y naturaleza sistemática, el desarrollo de esta área de estudio: el International Corpus of Learner English (ICLE) –cuya segunda versión fue lanzada en 2009– y el International Database of Spoken English (LINDSEI) –publicado en 2010–. Las cuestiones metodológicas para analizar estas

[1] Los términos *segunda lengua* y *lengua extranjera* se emplean en este trabajo indistintamente para referirse a la lengua meta del aprendiente.

[2] Esta terminología en inglés se consolida una vez que se compilan los primeros corpus de aprendientes informatizados, lo que implica que el corpus se almacena y que se accede a él electrónicamente, y que se cuenta con herramientas y *software* disponible para realizar búsquedas avanzadas. Estas cuestiones, que se aplican a la Investigación de Corpus de Aprendientes, no se asimilan necesariamente al Análisis de la Interlengua basado en corpus, terminología extendida en la tradición investigadora del español como lengua extranjera (ELE) para referirse también a los estudios basados en una colección de muestras de lengua natural, no necesariamente informatizada, que ha sido compilada para un estudio lingüístico; este es el caso de conocidos trabajos como el de Fernández (1997). La traducción al español de los términos y siglas en inglés es nuestra.

nuevas bases de datos informatizadas han sido también una de las principales preocupaciones de Granger; al mismo tiempo que se desarrollaba el corpus ICLE, se diseñaba un marco de investigación del corpus de aprendientes, conocido como Análisis Contrastivo de la Interlengua o ACI[3] (del inglés *Contrastive Interlanguage Analysis* o CIA), recientemente reivindicado por Granger (véase Granger 2015). Este método, a diferencia de los anteriores métodos de análisis de la lengua del aprendiente –análisis contrastivo (AC) y análisis de errores (AE)–, trata la lengua del aprendiente –esto es, la IL– como una entidad autónoma e independiente, y por su diseño comparativo posibilita el descubrimiento de muchos rasgos distintivos de la IL a través de las comparaciones sistemáticas cuantitativas y cualitativas. Por un lado, permite dichas comparaciones entre dos variedades de la misma IL (p. ej. el análisis del español L2 de un grupo de francófonos *vs*. el del español L2 de un grupo de italófonos permite distinguir los rasgos de la IL que dependen de la lengua materna de los rasgos que son comunes a todos los aprendientes); las variedades de la IL subrayan la gran variabilidad en la lengua del aprendiente, que habría de ser tenida en cuenta en la Investigación de Corpus de Aprendientes; así, existen variedades determinadas por el nivel de dominio (p. ej. la comparación entre un corpus de nivel inicial y un corpus de nivel intermedio), por la motivación, por la edad, por el efecto del entorno de aprendizaje, por el efecto de tareas, etcétera. Por otro lado, pueden establecerse comparaciones entre una variedad de la IL y una variedad nativa o una variedad experta, normalmente nativa[4] (L2 *vs*. L1). De este tipo de contraste que sitúa el corpus del aprendiente frente a un corpus de producción experta y/o nativa se obtienen las semejanzas y las diferencias –referidas estas últimas a los usos de los HNN sobre e infrarrepresentados– entre el español de ambos grupos en una situación comparable y, por lo tanto, los rasgos lingüísticos que, en principio, los HNN habrán de adquirir o emplear en otros estadios de la adquisición si desean alcanzar la naturalidad y expresividad propias del experto y/o nativo. Este enfoque comparativo entre la L2 y la L1 ha recibido críticas a lo largo de estos años por parte de investigadores como Cook (1999) o Hunston (2002), entre otros, que lamentan que la lengua del aprendiente no sea tratada como un sistema lingüístico independiente, en el sentido de que si se establece que los aprendientes toman las normas de los nativos como meta, no pueden ser concebidos más que como nativos fracasados. Granger (2015: 13-14) hace frente a estas críticas –basadas en el rechazo a la comparación de la interlengua con la producción nativa– con sólidos argumentos partiendo de diferentes autores: en el proceso de adquisición, el aprendiente hace comparaciones cognitivas entre las dos lenguas

[3] La traducción al español de los términos y siglas en inglés es nuestra.

[4] En línea con el reciente planteamiento de Granger (2015: 16), incorporamos la noción "variedad experta" como añadidura de "variedad nativa", ya que, en función del objeto del análisis, la producción nativa no experta puede constituir un buen corpus de control pero no por ello un modelo de lengua para el aprendiente; p. ej., a la hora de analizar ciertas estructuras propias de las reseñas académicas conviene comparar el corpus conformado por reseñas académicas escritas por aprendientes de español universitarios con un corpus de control que esté compuesto por reseñas académicas redactadas por autores profesionales de reseñas (expertos, ya sean nativos o no nativos) y no por reseñas de estudiantes nativos universitarios de grado que no están necesariamente familiarizados con este género textual; en este último caso, estaríamos ante un corpus de control nativo no experto y mal seleccionado, de acuerdo con el objetivo de la investigación. En cambio, si el objetivo es descubrir el uso de conectores discursivos frecuentes en la producción escrita no nativa, basta con contrastar esta última con un corpus nativo comparable –convertido en una herramienta más útil para realizar comparaciones cuantitativas que un extenso corpus general del tipo CREA, CORPES XXI o Corpus del español– no necesariamente experto, como podría ser el CEDEL2, que está formado por un corpus de redacciones de aprendientes de español, fundamentalmente universitarios, y un corpus comparable de textos redactados por nativos universitarios que son, en su mayoría, escritores inexpertos. Así, debemos asegurarnos de que el corpus de control utilizado representa el estado de lengua que queremos analizar y que solo se contrastan los componentes del corpus que han sido diseñados para ser comparados.

continuamente, de ahí que un análisis basado en normas externas (las de la lengua meta) pueda ser psicolingüísticamente válido (Ellis y Barkhuizen 2005, citado en Granger 2015); en esta misma línea, no es una coincidencia que todo estudio de ASL integre una noción de la interlengua tras la que acecha siempre la lengua meta (Sung Park 2004, citado en Granger 2015), por lo que determinar los aspectos de la IL que presentan más probabilidades de divergir de la lengua meta es un objetivo legítimo (Lardiere 2003, citado en Granger 2015). Así pues, no hay que dejar de comparar la L1 con la L2, sino dejar de extraer ciertas conclusiones problemáticas cuando la actuación de los HNN no coincide exactamente con la de los hablantes nativos (HN), como señala White (2003: 27).

Por otro lado, los usos erróneos de los HNN son difíciles de identificar a través de este enfoque contrastivo, por lo que debe completarse con el AE o el Análisis de Errores asistido por ordenador (Granger 2008; del inglés *Computer-aided Error Analysis*; para una aplicación de este método al español, véase Campillos Llanos 2014); este último método adolece de las mismas limitaciones que el AE, como son la imposibilidad de describir toda la competencia lingüística del aprendiente –ya que se atiende únicamente al error– o de detectar las estructuras que los HNN tienden a evitar por su dificultad o por su desconocimiento. El ACI, en cambio, sí permite examinar la competencia lingüística más allá del error y, por lo tanto, rescatar los usos infrarrepresentados en la producción no nativa. Así pues, ambos métodos, ACI y Análisis de Errores (asistido o no por ordenador) se complementan en la tarea de proporcionar mejores descripciones de la interlengua. Y si, a su vez, al método del ACI le aplicamos el doble enfoque descrito más arriba (L1 *vs*. L2 y L2 *vs*. L2), obtenemos una metodología ideal para comprender mejor los mecanismos de adquisición de la L2 y para diseñar herramientas y métodos de enseñanza más eficaces. Todo ello ha convertido al ACI en un método muy fructífero en los estudios de la interlengua del inglés. Prueba de ello es la gran cantidad de estudios de interlengua contrastivos originados a partir del ICLE[5], motivo por el que este corpus ha acabado inspirando la creación de corpus similares en otras lenguas, como es el caso del Corpus Escrito del Español como Segunda Lengua (CEDEL2; Lozano 2009; Lozano y Mendikoetxea 2013), uno de los más empleados en la investigación de la interlengua del español por su calidad de diseño.

Pese a que, como se acaba de señalar, ha habido un importante desarrollo en el terreno de los estudios de la lengua del aprendiente desde los primeros análisis contrastivos –entre la lengua materna del aprendiente y la lengua meta– hasta el análisis de la IL –producida de manera natural o a través de procedimientos experimentales– en la investigación de la ASL, el aspecto colectivo de la naturaleza de la IL del español todavía no ha sido suficientemente estudiado, cuando debería ser una parcela fundamental de la investigación de la ASL. El objeto de la investigación sobre la IL ha sido más la producción lingüística de HNN individuales –o pequeños grupos de varios individuos– que la de grupos numerosos con el mismo perfil[6]; no obstante, el interés

[5] La página *web* del proyecto ICLE contiene un apartado que recoge las publicaciones que se sirven de esta base de datos comprendida por producciones de aprendientes de inglés de diversas lenguas maternas y producciones de nativos en contextos comparables: <http://www.uclouvain.be/en-169926.html>.

[6] El estudio de Fernández (1997), pionero en los estudios de la interlengua del español, constituye una excepción a la anterior afirmación, pues se basa en un corpus escrito relativamente extenso –si se tiene en cuenta que a finales de los 90 todavía no existían los corpus de aprendientes de español informatizados–. Fernández comparara la IL de cuatro grupos de lengua materna distintos en tres estadios de evolución de su IL, por lo que, por su diseño comparativo, este estudio en el ámbito del ELE podemos considerarlo el germen del ACI

de investigadores y profesores debiera consistir en conocer los rasgos de la interlengua de un gran grupo homogéneo de HNN: esos rasgos son generalizables a todos los hablantes con un mismo perfil, es decir, no solo son aplicables al grupo analizado, por lo que las descripciones y explicaciones lingüísticas obtenidas por los investigadores son muy abarcadoras; y los profesores, si conocen los rasgos colectivos de un gran grupo, pueden asegurarse de que las necesidades cubiertas en el aula son las sentidas y compartidas por la mayoría de los aprendientes. Así, es necesario examinar el español producido por un gran grupo homogéneo más que por el de un individuo o varios[7]. Precisamente por todo ello, la investigación de la interlengua por medio del ACI basado en corpus es muy útil, ya que, por un lado, permite descubrir –a través de su función diagnóstica– los rasgos de la interlengua del español colectivo por medio de varias técnicas y procedimientos combinados, como son las relaciones de frecuencias y el test de significatividad, a los que se suma la aplicación de la metodología del AE; y, por otro lado, ayuda a encontrar –a través de la función evaluativa– indicadores de una actuación de dominio alto o bajo en el español no nativo.

Dedicamos los siguientes apartados al análisis de estas funciones por medio de las cuales los estudios que emplean la metodología del ACI contribuyen a la investigación de la interlengua. Para ello nos centramos fundamentalmente en el enfoque contrastivo del tipo L1 *vs*. L2.

2. La función diagnóstica

La función diagnóstica nos permite descubrir qué usan y qué no usan los HNN en comparación con el discurso nativo o experto, normalmente nativo. Las diferencias cuantitativas encontradas son examinadas posteriormente en un análisis cualitativo, lo que nos ofrece la posibilidad de conocer los rasgos de la interlengua del español colectivo y diagnosticar las posibles necesidades de los HNN, toda vez que, como señala Granger (2015), se hayan aplicado los filtros necesarios (objetivos de enseñanza, necesidades de los aprendientes, "enseñabilidad"[8]) que ayudan a determinar la conveniencia de actuar o no sobre esas diferencias cuantitativas reveladas por el ACI; por ejemplo, utilizar significativamente más los conectores discursivos más frecuentes (*pero*, *y*, *porque*) puede ser una buena estrategia en los primeros estadios de adquisición de una lengua, pero en el nivel avanzado puede implicar una falta de riqueza léxica o un desconocimiento de conectores más

basado en corpus –del tipo L2 *vs*. L2–, si bien este método de análisis, desde su aparición (1996) se viene aplicando únicamente a los corpus de aprendientes informatizados. Vázquez (1991), Santos Gargallo (1993) o Penadés (1999) son otras conocidas investigaciones de la lengua del aprendiente de español basadas también en una importante cantidad de datos, pero, en realidad, se centran en el AE, dado que los usos acertados o los sobre e infrarrepresentados no se consideran. El grueso del análisis de Fernández también recae en el estudio del error; en cambio, en estudios posteriores, tras la consolidación del ACI en el ámbito de la ASL, se está comenzando a llevar a cabo un análisis de la IL más abarcador que incluye de manera equilibrada el análisis del acierto y de todo tipo de desviaciones en la L2 –en relación con la L1 (error, sobrerrepresentación e infrarrepresentación)–; para un análisis de este tipo aplicado a la interlengua del español, véase Sánchez Rufat (2015).

[7] En <http://wdb.ugr.es/~cristoballozano/?page_id=64> se recogen las publicaciones que se sirven del CEDEL2.

[8] Véase el apartado "3. La función evaluativa" para una discusión sobre este concepto.

específicos (*no obstante*, *asimismo*, *ya que*); este último caso debería implicar atención en el aula, mientras que la primera situación no conllevaría una actuación docente.

Frecuentemente, los análisis de interlengua contrastivos se refieren a la divergencia entre las frecuencias en la lengua del aprendiente y la lengua nativa con los términos *sobreutilización* e *infrautilización* (del inglés *overuse* y *underuse*), terminología muy extendida en el ámbito de los estudios contrastivos de interlengua desde la publicación pionera de Granger (1998), *Learner English on Computer*. Ahora bien, como observan Guo (2006: 175) o Hasselgard y Johansson (2011: 55), entre otros, la interpretación de estos conceptos puede inducir a juicios erróneos, dado que se utilizan normalmente para indicar que cierto elemento es producido por los HNN de manera diferente –y se infiere que, por tanto, equivocada– a como lo hacen los HN:

> when people talk about 'overuse' or 'underuse' for a particular item, they imply that learners are using such an item wrongly. When people say a particular word is 'overused' by learners, they are implying that on some of the occasions it should no be used. By the same token, when people talk about the 'underuse' of the word, what they are implying is that learners do not use the word when they should use it (Guo 2006: 175).

Como ya se ha señalado, que los HNN utilicen un elemento con más frecuencia que los HN no indica necesariamente que ese elemento esté mal empleado o que deba emplearse menos, ya que puede tratarse de una estrategia que ayuda a los aprendientes a comunicarse o aprender de manera más efectiva (Granger 2015). No obstante, las diferencias cuantitativas sí que debieran suponer un punto de partida para un análisis posterior. Conviene, por lo tanto, matizar estos dos conceptos sobregeneralizadores –infrautilización y sobreutilización– que no implican necesariamente juicios cualitativos acerca de la actuación en la interlengua. De esta manera, en este trabajo se aboga por una descripción más rigurosa de la lengua del aprendiente por medio del empleo de los ocho esquemas de relación de frecuencias y de la aplicación del test de significatividad Chi-cuadrado[9]. La siguiente tabla, adaptada de Guo (2006: 176), nos permite, por un lado, establecer qué usos de los analizados son frecuentes y cuáles son infrecuentes en los HN y en los HNN; y, por otro lado, presenta una especificación a partir de los dos conceptos anteriores que contribuye a describir la lengua del aprendiente con mayor precisión.

Corpus de HNN	Corpus de HN	Interpretación posible
1. Alta frecuencia de X[10]	Alta frecuencia de X	X representa la parte de la lengua de los HNN más ampliamente compartida[11].
2. Alta frecuencia de X	Baja frecuencia de X	A menos que la diferencia se deba a una variación en el tema del texto –oral o escrito– (diferentes temas producen diferentes palabras nucleares), y si no se refiere a una buena estrategia de aprendizaje, existe una

[9] Para un estudio basado en estas herramientas de análisis combinadas aplicadas al uso del verbo *dar*, véase Sánchez Rufat 2015.

[10] "X" puede tratarse de un elemento perteneciente a cualquiera de los niveles lingüísticos del español o de sus interfaces, como el léxico, sintáctico, discursivo, léxico-sintáctico o sintáctico-discursivo, o incluso pragmático.

[11] Como se especifica más abajo, una frecuencia similar no garantiza semejanzas en el uso detallado, por lo que siempre se requiere realizar un análisis cualitativo de los datos.

		sobreutilización de X por desconocimiento del uso preferente en los HN. X puede ser un legado de los profesores de español. Si profesores y autores de materiales conocen qué palabras son usadas normalmente con demasiada frecuencia entre los HNN, pueden preparar materiales con las alternativas apropiadas registradas en los corpus nativos.
3. Baja frecuencia de X	Alta frecuencia de X	A menos que la diferencia se deba a una variación en el tema del texto, indica una infrautilización de X por desconocimiento del uso preferente en los HN y que, por tanto, otros elementos se están usando inadecuadamente en su lugar. Posiblemente las ocurrencias de X en los HNN en este esquema se correspondan con textos de un nivel de dominio de español más alto[12]. Esta área debe ser objeto de trabajo para los HNN si quieren progresar.
4. Baja frecuencia de X	Baja frecuencia de X	X no se usa con frecuencia en español en general o en este tipo de textos, por lo que no se dan suficientes oportunidades de uso para los HNN; posiblemente la mayoría de los HNN no son conscientes de este uso en español, por lo que existe la posibilidad de que las ocurrencias de X en los HNN en este esquema se correspondan con textos de un nivel de dominio de español más alto[13].
5. No casos de X	Baja frecuencia de X	La diferencia se debe al tema del texto o es un caso de infrautilización; al igual que en el esquema anterior, posiblemente X no sea suficientemente conocido por los HNN, quienes estarán usando en su lugar otras expresiones posiblemente inadecuadas (opciones no preferentes). Esta área debe ser objeto de trabajo para los HNN si quieren progresar.
6. Baja frecuencia de X	No casos de X	La diferencia se debe al tema del texto o hay una inadecuación por sobreutilización al usar algo que los HN no usan –lo esperable es que suceda lo contrario–; desconocen el uso preferente de los HN en ese contexto. Posiblemente no manejan la estructura o palabra que utilizan –ya que los HN no la usan–, por lo que puede contener errores.
7. No casos de X	Alta frecuencia de X	Si esta divergencia no se debe al tema del texto, sugiere una falta de dominio por parte de los HNN. Esta área debe ser objeto de trabajo para los HNN si quieren progresar. Si los HNN utilizaran los elementos más usados por los HN, ganarían en naturalidad[14].
8. Alta frecuencia de X	No casos de X	Si no se debe al tema del texto, y el uso es correcto, los HNN deben considerar otros modos de expresar la misma idea y que sean utilizados por HN.

Tabla1. Relación entre los índices de frecuencia nativos y no nativos

[12] Sobre la función evaluativa del ACI se reflexiona en el apartado 3.

[13] Ídem.

[14] A menudo, se dice de la producción no nativa que es bastante correcta gramaticalmente, pero que adolece de una falta de naturalidad, lo cual no es fácil de descubrir si no aplicamos esta metodología y si, por el contrario, realizamos únicamente un análisis de los errores.

Para poder interpretar los ocho tipos de relaciones de frecuencia entre los dos corpus que figuran en la Tabla 1, es necesario fijar primero los valores de las nociones *alta frecuencia* y *baja frecuencia*, que aparecen en la columna de la izquierda. Para un corpus que contiene en torno a medio millón de palabras, un porcentaje del 5 % o más se considerará alta frecuencia, y uno menor que el 5 % se considerará baja frecuencia. Aunque este límite es arbitrario, está basado en nuestra experiencia de análisis a través del CEDEL2 y coincide con el marcado por Guo (2006: 182), quien trabaja con un corpus de tamaño muy similar al CEDEL2, el College Learner English Corpus (COLEC).

Los tipos de relaciones de frecuencias 2, 3, 5, 6, 7 y 8 revelan las áreas en las que se producen las desviaciones de frecuencia entre la lengua nativa o experta, y la no nativa. Aparentemente, los tipos de relaciones de frecuencia 1 y 4 son de semejanza, pero, como se ha señalado, la semejanza no tiene por qué indicar un uso detallado similar, pues la 4 puede deberse a desconocimiento de los aprendientes, y tanto el esquema 1 como el 4 pueden esconder diferencias estadísticamente significativas en el índice de frecuencia –pese a que estemos refiriéndonos a una alta frecuencia de un ítem en ambos grupos o a una baja–. Por ejemplo, como se desprende del análisis del verbo *dar* en Sánchez Rufat (2015), los nativos utilizan la estructura ditransitiva con una frecuencia del 30,6 %, como en *Me da pena que mis hijos no puedan vivir esa experiencia*, y los HNN la emplean con una frecuencia del 48,4 %, como en *Por fin, me dio el billete*; estas frecuencias son altas (superan el 5 %), pero se diferencian en un 18 %, lo que es estadísticamente significativo ($p < 0,05$). Esta sobreutilización conlleva además la infrautilización de otras estructuras; por ejemplo, los HN prefieren el uso de un verbo en lugar de emplear el sustantivo derivado del verbo en coaparición con el verbo *dar*, como en *En Año Nuevo Emery y yo nos dimos besos a media noche porque es la tradición*, en lugar de *nos besamos*; estas dos estructuras no son intercambiables en todos los contextos (véase Sánchez Rufat 2015 para una discusión profunda al respecto).

En principio, desde una perspectiva comparativa entre los dos corpus, cuanto mayor sea la frecuencia de un ítem en el corpus nativo, más certeza habrá en cuanto a la significatividad de la ausencia o presencia del ítem en el corpus de aprendientes. De igual modo, cuanto menor sea la frecuencia del ítem en el corpus nativo, menos seguridad tendremos para referirnos a la significatividad de su presencia o ausencia entre los aprendientes. En relación con ello, el esquema 1 representa un área de madurez comparativa del español del aprendiente; los esquemas 3 y el 7 diagnostican mejor los problemas del aprendiente, ya que revelan las áreas en las que el español nativo presenta alta frecuencia de uso en X y el español del aprendiente se desvía con fuerza de aquel –siempre y cuando la divergencia no se deba al tema del texto–; los esquemas 2 y 8 también suponen un buen diagnóstico del español colectivo, ya que los rasgos de alta frecuencia presentes en los HNN representan lo popular y homogéneo del español colectivo de los HNN (Guo 2006: 178-179); esto es, dadas las mismas tareas (de narración) y bajo las mismas circunstancias, otros aprendientes con el mismo perfil probablemente producirán las mismas construcciones o elementos.

Pese a que esta perspectiva de interpretación de la relación entre las frecuencias alta y baja permite describir el español colectivo de los HNN con mayor detalle frente a lo que supondría limitarse a la sobreutilización o infrautilización de un elemento, no nos permite conocer con exactitud si las diferencias son estadísticamente significativas; así lo señalan McEnery y Hardie (2012: 51):

> To better understand and characterise the frequency data arising from a corpus, corpus linguistics appeal to statistical measures which allow them to shift from simply describing what they see to testing the significance of any differences observed. Most things that we want to measure are subject to a certain amount of random fluctuation. We can use significance tests to assess how likely it is that a particular result is a coincidence, due simply to chance.

De este modo, se debe aplicar siempre que sea posible un test de significatividad, como el del Chi-cuadrado, a las diferencias entre las frecuencias, con un 5 % como nivel crítico de significatividad estadística ($p < 0,05$). Como la mayoría de los aspectos que medimos están sujetos a cierto grado de fluctuación azarosa, por medio de este test es posible averiguar si los resultados obtenidos son, por tanto, estadísticamente significativos. Ello sucede cuando hay un 95 % de posibilidades de que nuestro resultado no se produzca por simple coincidencia.

Por otro lado, la frecuencia en el corpus de HNN solo revela cuántas veces ocurre un elemento determinado en el corpus. Aunque puede indicar cierto grado de adquisición en la producción del español por los HNN, no refleja necesariamente dominio, pues la lengua del aprendiente está cargada de rasgos problemáticos y expresiones artificiales. En palabras de Guo (2006: 183-84), "Large frequency in a particular pattern (or phrase) indicates only that this pattern (or phrase) is used fairly often by the group of writers. It has nothing to say about its appropriateness *per se*". Tras la frecuencia a menudo se ocultan los errores.

Ya se ha señalado (apartado 1) que la función de diagnóstico del ACI puede verse optimizada si se aplica a los datos un análisis del error. Con el objetivo de ofrecer un análisis del error lo más detallado posible, que contribuya a ofrecer una descripción precisa de la lengua del aprendiente y que, al mismo tiempo, pueda ser utilizado en la enseñanza de ELE, conviene presentar una propuesta tipológica del error diseñada expresamente para la investigación en cuestión, lo cual requiere, en primer lugar, fijar algún tipo de norma prescriptiva a partir de la cual analizar la interlengua y poder determinar, de esta manera, la adecuación de las muestras lingüísticas a la variedad del español seleccionada como modelo[15].

El estudio del error puede realizarse manualmente o pueden utilizarse las técnicas de la lingüística de corpus a través de la metodología del Análisis de Errores asistido por ordenador. En ambos casos, el tratamiento del error suele estar originado a partir del marco del AE propuesto por Corder (1971), que consiste en identificar, describir, clasificar y explicar los errores, si bien la metodología del Análisis de Errores asistido por ordenador conlleva también una etiquetación del error de acuerdo con su descripción y explicación; la propuesta taxonómica del error se suele configurar una vez que los errores son recopilados, y no antes, para no influir en el análisis con los tipos de errores que previsiblemente nos podríamos encontrar.

En suma, por todo lo argumentado anteriormente, la función diagnóstica del ACI debe ser considerada en toda investigación de corpus de aprendientes, ya que es la que nos permite conocer con rigor los rasgos de la interlengua del español colectivo y diagnosticar las necesidades de los HNN. Y ello se consigue por medio de la aplicación de varias técnicas y procedimientos combinados –los cuales se complementan entre sí–, a saber: 1) los ocho esquemas de la relación de frecuencias –con las respectivas asunciones basadas en la intuición– (adaptado de Guo 2006); 2) el test de significatividad Chi-cuadrado, aplicado a las diferencias cuantitativas

[15] En Sánchez Rufat y Jiménez Calderón (2013) se discute en profundidad la cuestión de la norma en los análisis de la interlengua.

obtenidas; y, fuera de los márgenes del ACI pero vinculado a sus contribuciones, 3) el análisis pormenorizado de las concordancias en el corpus no nativo, que da cuenta de los aciertos y errores por medio de un examen realizado a partir de una tipología del error que debe aspirar a ser consistente, informativa, flexible y reusable (cualidades destacadas por Granger 2003), y que permite identificar, localizar, describir y explicar los errores (basada en el tratamiento del error propuesto por Corder 1971).

Con un conocimiento de los errores y aciertos, y de las semejanzas y diferencias de uso entre el español de los HN y el de los HNN, podemos comprender y describir mejor el estadio de interlengua en el que se encuentran los HNN, y diagnosticar sus necesidades.

3. La función evaluativa

La segunda de las funciones que tiene un corpus de HNN es la evaluativa, que permite encontrar indicadores de una actuación de nivel de dominio alto y bajo en el español colectivo (Guo 2006: 231). El hecho de que los HNN no utilicen, utilicen con muy poca frecuencia o utilicen mal los elementos (ya sean léxicos, sintácticos, discursivos o pragmáticos) más usados por los nativos o expertos, indicaría que el español colectivo de los HNN no muestra un nivel de dominio lingüístico muy alto. De igual modo, el hecho de que los HNN no utilicen o utilicen significativamente menos los elementos poco frecuentes entre los HN (siempre y cuando estos usos nativos no se correspondan con usos individuales) parece reflejar un dominio colectivo del español relativamente bajo. De esta manera, la hipótesis resultante es que los grupos que no producen adecuadamente los usos frecuentes de los HN posiblemente estén en un estadio de adquisición más temprano, esto es, su nivel será más bajo –reflejaría un español menos amplio en la escritura y, por lo tanto, menos expresivo– que el de aquellos grupos que sí los producen correctamente.

De igual modo que el análisis de corpus nos permite evaluar el grado de dominio del español colectivo, la alta o baja frecuencia de un elemento determinado en el corpus nos puede servir para evaluar a aprendientes individuales. Como sugiere Guo (2006: 232), aquellos HNN que no producen o no producen bien usos comunes o frecuentes entre los HN estarían en un estadio de adquisición inferior, por lo que su nivel se correspondería con un dominio lingüístico general más bajo que el de aquellos aprendientes que sí los producen y lo hacen correctamente. Parece improbable que aquellos que no usan o no usan bien estos elementos comunes y frecuentes entre los HN obtuvieran una puntuación alta en la calificación del escrito que constituye la muestra del corpus. Asimismo, aquellos HNN que utilizan bien los elementos que no son muy frecuentes entre los HN mostrarían en el escrito un nivel de dominio más alto, por encima de la media, lo cual no implica que aquellos que los usen, pero mal, no tengan necesariamente un dominio alto.

Para poder contrastar estas intuiciones y, por tanto, aceptar esta función evaluativa del ACI, convendría calificar o acceder a la calificación que cada uno de los participantes del corpus hubiera obtenido en su escrito; si tras evaluar el texto en cuestión –para lo cual intervienen múltiples factores, como la coherencia y la cohesión,

la precisión y amplitud léxicas, el conocimiento ortográfico, gramatical, pragmático y sociolingüístico, entre otros– se comprueba que existe una correlación entre el uso adecuado de elementos usados por los HN con una frecuencia baja y el nivel general del texto, o entre el uso inadecuado de elementos usados con una frecuencia alta por los HN y el nivel general del texto, podría hablarse de una importante aportación al ámbito de la evaluación en ELE. Ilustramos esta función con el siguiente ejemplo: el análisis del verbo *dar* en Sánchez Rufat (2015), basado en el CEDEL2, revela que los HNN de nivel avanzado utilizan significativamente menos y en menor variedad que los HN (12,2% frente a 3,7%) ciertas construcciones pronominales y preposicionales en las que participa este verbo, como las siguientes: *darse* + SN, con el significado de "suceder o existir"; *dar por* + infinitivo y adjetivo ("acción inesperada" y "considerar"), *dar con* + SN ("encontrar") y *dar a* + infinitivo con el sentido de "efectuar acciones que denotan conocimiento o difusión de informaciones" o "percepción". Estas diferencias en la frecuencia de uso supone un foco de desviaciones, y, aunque no son estructuras tan frecuentes entre los nativos como otras (p. ej., las transitivas), son naturales, por lo que, a nuestro juicio, debieran producirse en los textos de nivel avanzado si se pretende que la actuación no nativa sea tan expresiva como la nativa. Partiendo de estos datos, nuestra hipótesis es que el grupo de aprendientes que conforma el corpus de nivel avanzado del CEDEL2 tiene un nivel de dominio relativamente bajo. Para confirmar o refutar la hipótesis anterior habría que evaluar las narraciones de todos los alumnos para determinar si el resultado es bajo para tratarse de un nivel avanzado; por otro lado, habría que comprobar si los individuos del grupo que hacen un uso adecuado de estas construcciones, en general poco utilizadas por los HNN, obtienen una calificación general más alta en la redacción que aquellos que no las emplean o las usan mal. De obtener, como grupo, una calificación no muy alta en el contexto de un nivel avanzado, ocurriría que de un análisis cuantitativo y cualitativo de los datos por medio del ACI se pueden obtener indicadores de nivel de dominio alto y bajo.

En suma, mientras que la función diagnóstica del corpus de aprendientes, cuando este se compara con el corpus experto o nativo, es explícita, la evaluativa es, más bien, implícita o potencial. Dominar el uso de las estructuras y elementos que son frecuentes entre los nativos hará que la producción no nativa se vaya aproximando gradualmente hacia el español nativo. En esta tarea tanto los profesores de ELE como los creadores de materiales tienen mucho que aportar, por lo que se espera que extraigan también el máximo beneficio de estas funciones, por medio de las cuales, como se ha expuesto en este trabajo, las investigaciones basadas en el ACI no solo contribuyen a los estudios de la interlengua del español, sino que también ofrecen al profesor información muy valiosa acerca de lo que los aprendientes dominan, desconocen o evitan. Así pues, la identificación de las desviaciones en el uso pueden constituir el punto de partida de aplicaciones en la enseñanza de ELE; como señala McCarthy (1990: 87), "predicting what learners will need (…) is important in selecting what to teach". Conviene matizar que, de acuerdo con la investigación en ASL, la enseñanza no conlleva necesariamente la adquisición; según la Hipótesis de "Enseñabilidad" (del inglés *Teachability Hypothesis*; véase Pienemann y Kessler 2012), el desarrollo del aprendizaje de muchas estructuras de la lengua meta es impermeable a la enseñanza, debido a que cognitivamente los aprendientes siguen un proceso de adquisición natural o predeterminado, por lo que no están preparados para aprender ni usar adecuadamente estructuras cuya adquisición está programada psicolingüísticamente en estadios posteriores. En consecuencia, como observa Lozano (2015), conviene saber que algunas implicaciones didácticas pueden no concordar con la Hipótesis de "Enseñabilidad".

A través de estas funciones, los HNN también pueden conocer los problemas que tienen como grupo y como aprendientes individuales, lo que les ofrece la posibilidad de atajarlos.

4. Conclusiones

Durante la última década, como respuesta al creciente interés por la adquisición del español como LE, han proliferado los corpus informatizados de aprendientes de español que proporcionan una base empírica sólida de lengua natural, como son los corpus escritos CEDEL2, CORANE (Cestero Mancera y Penadés 2009 [creado en 2001]) o CAES (Instituto Cervantes 2014); los orales SPLLOC (Mitchell *et al*. 2008) o CORELE (Campillos Llanos 2014); y el oral y escrito LANGSNAP (Tracy-Ventura *et al*. 2013). A pesar de este desarrollo de los corpus, se puede decir que la investigación de corpus de aprendientes de español está aún en sus inicios, pues los estudios, aunque crecientes, son todavía escasos. Al mismo tiempo, autores como Sánchez Rufat (2015) o Lozano (2015) han destacado las limitaciones teóricas y metodológicas de estas primeras investigaciones y de los corpus en sí mismos. Más esfuerzos deben ir encaminados en esta dirección; ya se ha señalado en este trabajo que el éxito de la investigación en ASL depende de la validez y fiabilidad de los procedimientos de obtención y análisis de datos, por lo que hemos de asegurarnos de que los corpus de aprendientes son capaces de combinar un buen diseño (Sinclair 2005), buenas herramientas de búsqueda y anotaciones; de lo contrario, no será posible explorar todas las preguntas de investigación imaginables ni tampoco beneficiarse del potencial descriptivo y, sobre todo, explicativo de las dos funciones analizadas y propuestas en este trabajo para los estudios de interlengua de español.

Anna Sánchez Rufat
Universidad de Extremadura
annasanchezrufat@hotmail.com

AGRADECIMIENTOS
Quiero dejar constancia de mi agradecimiento a los revisores de la revista; sus comentarios, observaciones y sugerencias han resultado sumamente útiles para la versión final de este trabajo.

Referencias bibliográficas

Campillos Llanos, L. (2014): "A Spanish Oral Learner Corpus for Computer-Aided Error Analysis", *Corpora*, 9 (2), pp. 207–38.

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.

Cook, V. (1999): "Going beyond the native speaker in language teaching", *TESOL Quarterly* 33 (2), pp. 185–209.

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

Fernández, S. (1997): *Interlengua y análisis de errores*, Madrid: Edelsa.

Granger, S. (1996): "From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora", K. Aijmer *et al*. (eds.), *Languages in Contrast. Text-based crossed linguistic studies*, Lund: Lund University Press, pp. 37-51.

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

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

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

Granger, S. (2008): "Learner corpora", A. Lüdeling y M. Kytö (eds.), *Corpus Linguistics: An International Handbook* (vol. 1), Berlín: Mouton de Gruyter, pp. 259–75.

Granger, S. (2015): "Contrastive interlanguage analysis: A reappraisal", *International Journal of Learner Corpus Research*, 1 (1), pp. 7-24.

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: 09-08-2015].

Hasselgard, H. y Johansson, S. (2011): "Learner corpora and contrastive interlanguage analysis", F. Meunier *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),* [en línea] <http://galvan.usc.es/caes> [consulta: 09-08-2015].

Larsen-Freeman, D. y Long, M. (1994): *Introducción al estudio de la adquisición de segundas lenguas*. Madrid: Gredos.

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

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

Lozano, C. (2015): "Leaner corpora as a research tool for the investigation of lexical competence in L2 Spanish", *Journal of Spanish Language Teaching,* 2 (2) (número monográfico titulado *New Perspectives on the*

Acquisition and Teaching of Spanish Vocabulary/ Nuevas perspectivas sobre la adquisición y la enseñanza del vocabulario del español, coordinado por A. Sánchez Rufat y F. Jiménez Calderón).

McCarthy, M. (1990): *Vocabulary*, Oxford: Oxford University Press.

McEnery, T. y Hardie, A. (2012): *Corpus Linguistics*. Cambridge: University Press.

Mitchell, R., L. Domínguez, M. Arche, F. Myles y E. Marsden (2008): "SPLLOC: A new corpus for Spanish second language acquisition research", L. Roberts *et al.* (eds.), *EUROSLA Yearbook 8*, Amsterdam: John Benjamins, pp. 287–304.

Pastor Cesteros, S. (2004): Apr*endizaje de segundas lenguas. Lingüística aplicada a la enseñanza de idiomas*, Alicante: Universidad de Alicante.

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

Pienemann, M. y Kessler, J-U. (2012): "Processability Theory", S. M. Gass y A. Mackey (eds.), *The Routledge Handbook of Second Language Acquisition*, New York: Routledge, pp. 228–246

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

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

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

Tracy-Ventura, McManus, N. K. y Mitchell, R. (2013): "A Longitudinal Learner Corpus Investigation of Vocabulary Learning Before, During, and After Residence Abroad", conferencia presentada en el *Learner Corpus Research Conference*. Bergen (Noruega), 27–29 de septiembre de 2013.

Vázquez, G. (1991): *Análisis de errores y aprendizaje de español/lengua extranjera*, Frankfurt: Peter Lang.

White, L. (2003): "On the nature of interlanguage representation: Universal Grammar in the Second Language", J.D. Doughty y M. Long (eds.), The Handbook of Second Language Acquisition, Malden: Blackwell, pp. 19-42.