CORPUS PARA EL ESTUDIO DE LA INTERLENGUA DE ESTUDIANTES DE ESPAÑOL DE ORIGEN CHINO[1]

M.ª Ángeles Calero Fernández
Maribel Serrano Zapata
Universidad de Lleida, España

INTRODUCCIÓN

Ante la pujanza que está adquiriendo el español en China, es necesario plantearse cuáles son las necesidades específicas del alumnado sinohablante de forma diferente a como se ha venido haciendo. Hasta ahora se ha publicado en torno a un millar de trabajos dedicados a la enseñanza de ELE a estudiantes chinos. Es una cantidad muy notable, pero la mayoría son de difícil acceso y con poca o nula difusión. El grupo internacional de investigación SinoELE ha mejorado sustancialmente esta situación al hacer visibles estos estudios gracias a su base de datos bibliográfica, aunque no siempre esté activa. Son investigaciones de temática diversa, especialmente dedicadas a la gramática, la pronunciación y la expresión escrita, con un enfoque contrastivo (Blanco, 2019). También es diversa la metodología y los instrumentos de recogida de datos que se han utilizado. Pero lo que es común a los distintos trabajos es que la fuente de información o es el contraste de lenguas o es un número de casos demasiado reducido para extraer conclusiones realmente generalizables o para confirmar o refutar hipótesis de manera rigurosa.

En el año 2005 se creó en la Universidad de Lleida el *Diploma de Estudios Hispánicos*, que pretendía dar respuesta a la necesidad de formación de estudiantes chinos de movilidad que se trasladaban a esta universidad para realizar una estancia de seis meses o de un año buscando reforzar su aprendizaje de español. Para afrontar adecuadamente esta tarea, fue preciso determinar las características de estos alumnos en lo referente al nivel de dominio lingüístico,

1 Este trabajo se ha realizado dentro del proyecto *Elaboración y catalogación de un corpus de textos escritos en ELE producidos por estudiantes sinohablantes* que se está desarrollando en la Universidad de Lleida y que está financiado por el Ministerio de Ciencia, Innovación y Universidades (N.º de Referencia: FFI2016-80280-R) con la participación de Fondos Feder.

al grado de desarrollo de las destrezas y a las estrategias de aprendizaje. Desde esa fecha, alrededor de seiscientos estudiantes de español sinohablantes, procedentes de tres universidades chinas, han llenado las aulas de la Facultad de Letras de la Universidad de Lleida.

En estas circunstancias no es extraño que se plantearan preguntas como qué exponentes lingüísticos son los más difíciles de aprender por estos estudiantes y por qué, cuáles son los errores más frecuentes entre ellos y en qué niveles, de qué tipología son dichos errores, qué factores lingüísticos y extralingüísticos intervienen en su aparición y en la fosilización de algunos de ellos, qué eficacia tienen las diferentes formas de reparar el error, etc. Esta clase de preguntas de investigación solo puede responderse de manera científica si se parte de una cantidad sustanciosa de textos reales producidos por aprendices de ELE y recogidos con métodos propios de la lingüística del corpus. Por tal razón, un equipo de investigación de la Universidad de Lleida familiarizado con este tipo de estudiantes ha decidido construir un corpus de aprendices de español cuya lengua materna es el mandarín. Para ello, ha contado con la colaboración de la Universidad de Valencia, la Universidad de Salamanca, la *Tianjin Foreign University Studies* (TFSU), la *Binhai School of Foreign Affairs* (BSFA), la *Changzhou University* (CZU) y la *Soochow University* (SU). Una labor de estas características hubiera sido imposible de no haber contado con la financiación del Ministerio de Ciencia, Innovación y Universidades.

La construcción de este corpus tiene como objetivo último disponer de una base de datos consistente y rigurosa que permita analizar la interlengua española de aprendices sinohablantes y, de esta forma, concretar cuáles son los errores lingüísticos propios de cada estadio, cuáles son recurrentes y cuáles se fosilizan, para mejorar en la didáctica de ELE a este colectivo. En este trabajo daremos noticia de las características del corpus elaborado, de las posibilidades de consulta y explotación que tiene en estos momentos y de las que serán posibles en un futuro próximo.

1. LOS CORPUS DE APRENDICES

Los corpus de aprendices en ELE no son algo nuevo. Como puede observarse en la tabla 1, contamos con corpus orales y corpus escritos, constituidos algunos por muestras tomadas de estudiantes de diversas nacionalidades y otros, por aprendices que comparten su lengua materna.

Corpus orales	Corpus escritos
De diversas lenguas maternas: · *Spanish Learner Oral Corpus / Corpus Oral de Español como Lengua Extranjera* (CORELE) · Fono.Ele	**De diversas lenguas maternas:** · *Corpus para el análisis de errores de aprendices de ELE* (CORANE) · *Corpus de aprendices de español* (CAES) · *Corpus de Aprendices de Español como Lengua Extranjera* (CAELE) · *Anglia Polytechnic University Learner Spanish Corpus* (APU) · *Aprender a Escribir en Lovaina* (Aprescrilov)
De la misma lengua materna: · *Spanish Learner Language Oral Corpora* (SPLLOC 1 y 2) · *Spanish Corpus Proficiency Level Training (SPT)* · *Corpus Oral de Interlengua Bilingüe Español-Italiano* (CORINEI) · *Corpus oral de interlengua de aprendices de ELE serbios*	**De la misma lengua materna:** · *Corpus del español de los italianos* (CORESPI) · *Corpus Escrito del Español como L2* (CEDEL2) (inglés) · *Corpus de Aprendices Suecos de ELE* (SAELE) · *Tartu Learner Corpus of Spanish as a L3* (estonio) · *Japanese Learner Corpus of Spanish*
Corpus orales y escritos	
· LANGSNAP (aprendices anglófonos) · Corpus de Aprendices Taiwaneses de Español (CATE)	

Tabla 1 - Corpus de aprendices de ELE según el canal y la lengua materna

Los corpus escritos son más antiguos que los orales, sin duda por la complejidad que tiene la recogida y tratamiento de las muestras de habla viva, así como la exigencia de aparatos adecuados de grabación y reproducción del sonido. El más antiguo de los corpus escritos es el CORANE, cuyos textos se

recolectaron en el año 2000[2]. Le siguen el APU, con textos tomados antes de 2004, y el *Japanese Learner Corpus of Spanish,* antes de 2005. La primera tanda del Aprescrilov se reunió entre 2004 y 2005 (la segunda es de 2009 y 2010). El *Tartu Learner Corpus of Spanish as a L3* se inició en 2004 y finalizó en 2014. En fechas parecidas comenzó a construirse el CEDEL2, en 2006, y no ha dejado de crecer; tiene activo un aplicativo (http://learnercorpora.com/) que permite seguir recibiendo textos en español y en inglés. El SAELE se elaboró entre 2008 y 2010; el CAES, entre 2011 y 2013; el CAELE, entre 2014 y 2015; y el CORESPI estaba acabado en 2018. El propio CATE tuvo primero su corpus escrito, el *Corpus Escrito de Aprendices Taiwaneses de Español* (CEATE), entre 2005 y 2011, y, posteriormente, entre 2013 y 2017, el *Corpus Oral de Aprendices Taiwaneses de Español* (COATE)[3]. En cambio, no tenemos ningún corpus oral hasta 2006, el SPLLOC 1 y 2; le sigue el SPT, de 2007; el CORINEI, elaborado entre 2009 y 2014; el CORELE, construido entre 2010 y 2012; Fono.Ele, que recoge muestras desde 2010; y el *Corpus oral de interlengua de aprendices de ELE serbios*, desde 2011.

Alonso (2016) señala que la mayor parte de los corpus de aprendices de ELE son escritos, como acabamos de ver, transversales, con textos que desarrollan temas prefijados, sobre todo en género narrativo o argumentativo, redactados por estudiantes de español de distintos niveles de dominio de la lengua extranjera, por lo general adultos y anglófonos. Según esta descripción, no sería esperable encontrar corpus de aprendices chinos; sin embargo, los hay, y los relacionamos en la tabla 2.

2 Precisamente la fecha de recolección de este corpus de aprendices explique que sea consultable en un soporte físico, esto es, en un CD, y no a través de internet, como lo son buena parte de los corpus siguientes. El CORANE sigue el ejemplo de otros corpus coetáneos, como son los del PRESEEA, lo que supone una mejora con respecto a los corpus que hasta entonces se habían presentado en papel.

3 El LANGSNAP, que ya hemos visto que contempla muestras tanto orales como escritas, se acopió entre 2011 y 2013.

Corpus	Lenguas de origen	Tipología textual	Niveles	Volumen
CAES	Árabe, **mandarín**, francés, inglés, portugués y ruso	Extensa y diversa. Remeda las pruebas de los DELE	A1-C1	575 000 palabras
CEATE	No se especifica	Textos ensayísticos	No consta	340 000 palabras

Tabla 2 - Corpus escritos de aprendices de ELE sinohablantes y sus principales características

Como puede verse, el CAES recoge textos que no son exclusivos de aprendices chinos, sino de hasta seis lenguas maternas distintas. En total, según los datos ofrecidos en la web del CAES (http://galvan.usc.es/caes/tareas), las tareas procesadas provenientes de aprendices cuya lengua materna es el mandarín suponen un reducido, para nuestros intereses, 9,6% del total de la muestra de tareas que alimentan este corpus. En el caso del CEATE, al no especificarse la lengua de origen, hemos de suponer que, como mínimo, la mayoría de los textos procederán de estudiantes cuya primera lengua es el mandarín, pero, dada la situación sociolingüística de Taiwán, la lengua materna podría ser otra (especialmente, el taiwanés). Por otra parte, el CEATE resulta monocromático en la tipología textual, puesto que recoge únicamente textos ensayísticos; en cambio, el CAES tiene una diversidad de tipos textuales. También difieren en el volumen de palabras: el CEATE alcanza algo más de la tercera parte de la magnitud del CAES, pero no debemos olvidar que este último corpus ha contado con la participación de veintiocho centros del Instituto Cervantes y siete universidades, casi todas españolas.

Este panorama justifica la creación de un corpus de aprendices de ELE de procedencia china y hablantes de mandarín, que sea exclusivo, recoja un abanico de géneros del discurso, desplegado en diferentes niveles de dominio de la lengua y que contemple otros rasgos que puedan ser considerados importantes en la caracterización posterior de la interlengua española de sinohablantes.

2. EL *CORPUS DE INTERLENGUA ESPAÑOLA DE APRENDICES SINOHABLANTES* (CINEAS)

El *Corpus de Interlengua Española de Aprendices Sinohablantes* (CINEAS) es el resultado más importante del proyecto *Elaboración y catalogación de un corpus de textos escritos en ELE producidos por estudiantes sinohablantes* (N.º de Referencia: FFI2016-80280-R). Este proyecto, iniciado en diciembre de 2016 y acabado en diciembre de 2019, ha tenido, como objetivo general, construir un corpus digitalizado extenso y representativo del español utilizado por sinohablantes que permita realizar investigaciones rigurosas sobre la interlengua y el proceso de enseñanza-aprendizaje en este colectivo. Para alcanzar este objetivo general, se han planteado los siguientes objetivos específicos:

1. Determinar unos criterios de selección y catalogación que permitan construir una base de datos extensa y representativa del español utilizado por sinohablantes en el registro escrito.

2. Recoger muestras lingüísticas de sinohablantes siguiendo los criterios de selección convenidos y catalogarlas.

3. Establecer un conjunto de categorías que permitan localizar tipos de rasgos lingüísticos dentro del corpus lingüístico constituido.

4. Catalogar digitalmente los textos que formen parte del corpus lingüístico.

El conjunto de investigadoras e investigadores que han participado en este proyecto ha sido M.ª Ángeles Calero Fernández (UdL), F. Javier Terrado Pablo (UdL), Montse Noria Jové (UdL), M.ª Begoña Gómez Devís (UV), Moisés Selfa Sastre (UdL), Maribel Serrano Zapata (UdL), Roberto García González (UdL), Carmen Vanesa Álvarez Rosa (USal), Maite Miramón Escribano (UdL), así como Zhang Peng (*Tianjin Foreign Studies University* - TFSU), Zhang Bohua (*Binhai School of Foreign Affairs* - BSFA), Li Xiaoke (*Soochow University* - SU), Xu Shuo (*Changzhou University* - CZU), Li Xiang (TFSU), Chen Lei (TFSU), Guo Yuanchun (BSFA) y Liu Zhuoyu (BSFA).

El CINEAS es un conjunto de textos escritos producidos por estudiantes de español con diferentes niveles de dominio, en concreto, de A1 a C1, según los niveles establecidos en el *Marco Común Europeo de Referencia* (MCER, 2002), y que tienen el mandarín como lengua materna. Se trata de una herramienta informática a la que se accede a través de la dirección https://cineas.udl.cat/, que va a permitir investigaciones aplicadas de todo tipo. La principal que persigue el equipo que ha construido CINEAS es la de identificar dificultades de aprendizaje, los errores más comunes y su evolución en las distintas fases de la interlengua. Pero puede proporcionar, además, información sobre preferencias léxicas, sobre estrategias discursivas, sobre coherencia y cohesión, así como datos de carácter sociolingüístico y pragmático-cultural. También puede ser una fuente de ejemplos para trabajar en el aula de ELE.

Los textos recogidos pertenecen mayoritariamente[4] a cuatro géneros discursivos: descripción, narración, exposición y argumentación, porque se parte de la base de que cada uno de estos tipos textuales promueve unos exponentes lingüísticos diferentes, con lo que se alcanzaría una visión panorámica del uso de la lengua. Por ello, se han recolectado muestras de todos los géneros para cada nivel de referencia del MCER (2002) contemplado en el CINEAS. Estos textos han sido obtenidos en dos tipos de contextos educativos distintos, a saber, en el país de origen y en inmersión lingüística. Los informantes son estudiantes de la Carrera de Español (más o menos equivalente a nuestro grado de Filología Hispánica) o de Máster, pertenecientes a las universidades chinas que participan en el proyecto, o que se encuentran en situación de movilidad en las universidades españolas también participantes en el proyecto.

Los textos se han redactado a mano a partir de un estímulo que consistía en un tema, unas instrucciones y un listado de contenidos para facilitar cierta homogeneidad. Asimismo, se acotaba el número de palabras mínimo y máximo, en progresión ascendente a medida que se sube en nivel de dominio de la lengua, y siempre por encima de los márgenes que se establecen en las pruebas escritas de los *Diplomas de Español Lengua Extranjera* (DELE). Igualmente, se establecía un tiempo máximo para elaborar el texto. Los mismos informantes han ido entregando diferentes tareas a lo largo del curso y, en muchos casos,

4 Hay algunos textos que reproducen diálogos, pero son los menos.

a lo largo de varios cursos académicos. La recogida se ha hecho desde el curso académico 2016-2017, hasta el curso académico 2019-2020. Por consiguiente, estamos ante un corpus longitudinal.

Los textos han sido catalogados a partir de diferentes factores sociales: sexo o género, curso, nivel de referencia, universidad, nivel de conocimiento del inglés (según el *College English Test*[5]), estancias en España y duración de las mismas. También se han catalogado por el factor lingüístico género discursivo. Todos estos elementos se han convertido en metadatos que permiten filtrar las búsquedas que quieran practicarse en el CINEAS. La selección de estas categorías para catalogar los textos responde a la hipótesis de que todas ellas influyen, en mayor o menor medida, en el conocimiento y uso del español por parte de aprendices sinohablantes. Los estudios sociolingüísticos han demostrado que el factor género o sexo tiene una incidencia directa en el comportamiento lingüístico, así como también la tiene el tipo de centro de enseñanza (por ejemplo, público o privado, o si está situado en un núcleo urbano pequeño o grande). La relación entre la corrección o incorrección lingüísticas y el nivel de dominio de la lengua española es evidente, pero también es crucial en el uso más o menos correcto del español el nivel de conocimiento acreditado del inglés. El inglés resulta ser la primera lengua extranjera que se aprende en las aulas chinas y a ella está expuesto el alumnado desde la enseñanza primaria. Por su similitud con la lengua española (ambas son lenguas indoeuropeas), sirve, en muchos casos, de puente para el aprendizaje del español, con los beneficios y los perjuicios que esa estrategia puede acarrear.

El CINEAS contiene un anotador lingüístico diseñado por el propio equipo de investigación, el *TextAnnot*. El *TextAnnot* es una aplicación web desarrollada con Spring Boot[6] que funciona como una arquitectura de *software* que incluye un lenguaje de programación, bibliotecas, programas y otros elementos que sirven para construir y entrelazar los componentes de aquello que se ha creado, en este caso, el corpus y el anotador. El código fuente que se utiliza es abierto, de acceso libre y reutilizable. No precisa instalarse en ningún ordenador porque se accede a él a través de cualquier navegador (cf. Calero, Serrano y Gómez-Devís, 2020). Por defecto, el *TextAnnot* funciona, entre otras cosas, como un anotador

5 El *College English Test*, o CET, es una prueba de acreditación oficial de la lengua inglesa en China.

6 Puede consultarse y utilizarse desde https://github.com/rhizomik/TextAnnot.

de errores. Ofrece un listado de errores organizados en árbol, pertenecientes a todos los niveles de la lengua (fonético-fonológico, morfosintáctico, léxico y textual), con los que se puede ir etiquetando los textos que forman parte del corpus[7]. La interfaz diseñada hará posible múltiples consultas a las que se podrán aplicar distintos filtros.

3. ALGUNOS DATOS ESTADÍSTICOS PROVISIONALES

En el momento en que se redacta este texto, no pueden darse aún los datos definitivos del CINEAS, que no estarán listos hasta diciembre de 2019. Pero sí podemos ofrecer una aproximación a partir de los últimos textos procesados a principios de septiembre de 2019.

Los datos de la tabla 3 nos informan del volumen del CINEAS, superior en cantidad de palabras al CEATE, pero inferior al CAES. El número de muestras[8] ya introducidas en el CINEAS a inicios de septiembre, es 3905, muy próximo a las 3878 del CAES, pero aún serán muchas más cuando acabe todo el proceso. De la tabla 3 también se desprende que el género discursivo minoritario es la argumentación. No obstante, hay que advertir que, cuando se extrajeron estos datos provisionales, faltaban cargar los textos de B2 y C1 ya digitalizados en XML y en los que se practicaron más textos argumentativos.

unidades	**n.º**	**género discursivo**	**%**
textos	3905	descripción	25,5
palabras	435944	narración	33
		exposición	35,7
		argumentación	5,81

Tabla 3 - N.º de textos y de palabras y porcentaje de textos según el género discursivo en el CINEAS

7 Dejamos para otra ocasión detallar el sistema de etiquetado de errores que ha sido la base del *TextAnnot* en comparación con otros ya existente, aunque ya en Calero, Serrano y Gómez-Devís (2020) se describen otras herramientas informáticas que se están utilizando para anotar errores.

8 El CAES las llama *tareas*.

La tabla 4 nos da noticia de las características de las personas que han aportado textos al CINEAS. Se observa que los niveles de referencia más bajos son los más representados en el corpus[9], que la contribución de las universidades es desigual (depende notablemente de las dimensiones de cada una de ellas), que las informantes son mayoría aplastante, que las tres cuartas partes tienen acreditado, al menos, el nivel intermedio de inglés y que son muy pocas las personas que han disfrutado de una estancia en España.

nivel	%	Universidad	%	sexo	%	Nivel CET	%	Estancia	%
A1	31,9	Binhai	32,7	mujer	80,7	0	26,8	nunca	90,4
A2	32,1	Changzhou	26	hombre	19,3	4	43,1	un semestre	2,4
B1	26,6	Tianjin	21,3			6	30,1	un curso	4,9
B2	9,3	Soochow	10					dos cursos	2,3
C1	0,1	Lleida	8,7						
		Salamanca	1,3						

Tabla 4 - Porcentajes de muestras que cumplen las siguientes características

CONCLUSIONES

La lingüística del corpus ha venido a demostrar las bondades de trabajar con grandes cantidades de datos, por ello, es el momento de analizar el español como lengua extranjera a partir de un volumen de muestras lingüísticas suficiente para hacer fiables los resultados y para que sea posible replicar estudios. Tal cosa será factible si se dispone de bases de datos consistentes. Lamentablemente, hasta ahora no se disponía, entre los corpus de aprendices de ELE, de uno que sirviera para investigar ampliamente y con rigor el español aprendido por estudiantes chinos, atendiendo a ciertas variables lingüísticas y extralingüísticas. El CINEAS viene a cubrir ese vacío. Se trata de una base de datos consistente, elaborada con textos escritos por estudiantes sinohablantes

9 No se olvide la advertencia de que faltaban por cargar textos de B2 y C1, por lo que la fotografía cambiará sustancialmente en este punto.

de diferentes universidades de China que permanecen en su universidad o que están realizando una estancia de movilidad en España. Este corpus longitudinal va acompañado de un anotador de errores que facilitará identificar cuáles son los más frecuentes, cómo van evolucionando y si se fijan irremediablemente, intentando descubrir en qué medida contribuyen a su etiología un conjunto de rasgos como el sexo o género o el nivel de dominio del inglés, u otros condicionantes que puedan hacerse visibles gracias a los datos estadísticos que se podrán obtener. Uno de los valores del CINEAS es que es accesible a la comunidad científica. En la línea de las iniciativas europeas de investigación abierta (especialmente la que tiene financiación pública) se ha querido ofrecer la aplicación bajo una licencia abierta para fomentar su uso y futuro mantenimiento. Cualquiera puede contribuir al código y ayudar a mantenerlo, siempre que haga sus contribuciones también abiertas.

REFERENCIAS BIBLIOGRÁFICAS

Alonso Ramos, M. (2016). Spanish Learner Corpus Research. Achievements and challenges. En M. Alonso-Ramos (Ed.), *Spanish Learner Corpus Research. Current trends and future perspectives* (pp. 3-32). Amsterdam/Philadelphia: John Benjamins.

Blanco Pena, J. M. (2019). Estudios sobre la enseñanza de E/LE a sinohablantes: Estado cuantitativo de la cuestión. *SinoELE. Revista de Enseñanza de Español a Sinohablantes, 18*, 27-38.

Calero Fernández, M.ª Á., Serrano Zapata, M. y Gómez-Devís, M.ª B. (2020). Codificación y etiquetado en los corpus de aprendices y su aplicación didáctica: la propuesta del Corpus de Interlengua Española de Aprendices Sinohablantes (CINEAS). *E-AESLA, 6*, 206-222. Recuperado de https://cvc.cervantes.es/lengua/eaesla/pdf/06/15.pdf

Consejo de Europa (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Madrid: MECD/Anaya. Recuperado de http://cvc.cervantes.es/obref/marco