CODIFICACIÓN Y ETIQUETADO EN LOS CORPUS DE APRENDICES Y SU APLICACIÓN DIDÁCTICA: LA PROPUESTA DEL *CORPUS DE INTERLENGUA ESPAÑOLA DE APRENDICES SINOHABLANTES* (CINEAS)[1]

Mª Ángeles Calero Fernández
Maribel Serrano Zapata
Universidad de Lleida
Mª Begoña Gómez-Devís
Universidad de Valencia

RESUMEN

Los criterios con los que se diseña, se cataloga y se etiqueta un corpus lingüístico determinan las aplicaciones que dicho corpus podrá tener (Leech, 1993). En esta publicación se revisa la codificación y el etiquetado de distintos corpus de aprendices de español como lengua extranjera (ELE), así como algunos programas informáticos utilizados para ello, con el fin de evaluar el tipo de información de estos corpus que puede ayudar a caracterizar los estadios de la interlengua y a determinar los factores lingüísticos y extralingüísticos que intervienen en la formación, estructura y progreso de la interlengua.

Palabras clave: lingüística del corpus, corpus de aprendices, enseñanza de ELE

ABSTRACT

The criteria for designing, encoding and annotating a learner corpus determine its potential as a research tool and as a pedagogical resource. This paper reviews the criteria employed in coding and annotating a series of written learner corpora of Spanish. It also examines the possibilities these corpora offer for the following tasks: 1) identifying the features that characterize the different stages of the interlanguage of learner Spanish; 2) determining which linguistic and extralinguistic factors intervene in the formation, structure and progress of the interlanguage; and 3) determining which factors intervene in the teaching-learning process of a foreign language.

[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 Economía y Competitividad (Nº de Referencia: FFI2016-80280-R) con la participación de Fondos Feder.

Keywords: corpus linguistics, corpus of apprentices, teaching of ELE

1. INTRODUCCIÓN

La pujanza actual de las investigaciones basadas en corpus no se concibe sin el notable desarrollo de los corpus textuales en lengua inglesa, iniciado por el estructuralismo americano de la segunda mitad del siglo XX. A partir de entonces la evolución ha sido muy rápida pues los avances originados por la necesidad de estudiar la lengua en uso no solo han alcanzado al desarrollo de la lingüística como disciplina sino también a la informática, que ha facilitado cada vez más el manejo de ingentes cantidades de datos; asimismo han sido cada vez más las lenguas sobre las que se han ido construyendo corpus y estos han ido derivando en toda una tipología.

Una definición básica de corpus lingüístico fácilmente reconocible podría ser: conjunto de textos informatizados producidos en situaciones reales que se han seleccionado siguiendo una serie de criterios lingüísticos explícitos los cuales garantizan que dicho corpus pueda ser usado como muestra representativa de la lengua (EAGLES, 1996). Ciñéndonos exclusivamente a consideraciones didácticas, las competencias de los corpus son numerosas. Para el alumno, el corpus constituye una base sólida para identificar las estructuras lingüísticas más frecuentes en las producciones reales de los hablantes nativos de una lengua; le otorga la autonomía de elegir por sí mismo qué aprender, cómo aprenderlo y en qué orden, y, además, es un instrumento para encontrar respuestas a una tipología muy variada de dudas concretas y de profundizar en ellas por medio del acceso a amplios contextos reales. Al profesor no nativo, el corpus le permite basarse no exclusivamente en su intuición y en ejemplos elaborados *ad hoc*, sino en una fuente amplia y fiable de recursos lingüísticos, así como seleccionar un *input* suficiente y de calidad al que enfrentar a sus alumnos (Pérez, 2007: 11). Los corpus de aprendices son un tipo de corpus lingüístico que recoge muestras de habla (oral y/o escrita) producidas por hablantes no nativos; su finalidad es registrar la forma en que estos se expresan en la lengua extranjera y, en consecuencia, aporta al docente información muy valiosa para la planificación de la enseñanza.

Pero no se puede acceder a la información de un corpus si este no ha sido preparado para buscar datos de manera rápida, sistemática y automática. Además de digitalizarlo, es necesario codificarlo y etiquetarlo, y la forma en

que se realicen estas tres tareas condicionará indefectiblemente las aplicaciones que dicho corpus pueda tener y el tipo de explotación que pueda acometerse (Leech, 1993).

Nuestro trabajo va a centrarse en la forma en que han sido diseñados los corpus de aprendices y su efecto en la enseñanza-aprendizaje de la lengua extranjera. Los objetivos que perseguimos son: 1) describir la forma en que se ha codificado y etiquetado los corpus de aprendices de ELE a partir de los metadatos que se tienen en cuenta, 2) presentar los programas informáticos utilizados para anotarlos, incluido uno recientemente creado en la Universidad de Lleida, y 3) por último, evaluar si los metadatos y la anotación que se han empleado pueden ayudar a caracterizar las etapas de la interlengua y los elementos lingüísticos y extralingüísticos que intervienen.

Así pues, daremos primero un repaso a este tipo de corpus y a los metadatos que utilizan. Seguidamente veremos distintos anotadores. Finalmente, concluiremos la utilidad didáctica que tiene todo ello.

2. LOS CORPUS DE APRENDICES DE ELE

Los corpus de aprendices no son desconocidos en ELE. Varias autoras (Penadés 2005, Baralo 2010, Ainciburu 2012, Buyse, Fernández y Verveckken 2016) han señalado la necesidad de bancos de datos longitudinales de ELE en diferentes contextos de aprendizaje de la lengua que puedan servir de punto de referencia para planificar su enseñanza. Y en los últimos 15 años han ido surgiendo corpus que recogen textos orales o escritos producidos por estudiantes de español. La diversidad que presentan estos corpus hace imprescindible describirlos para que docentes e investigadores puedan conocer las posibilidades de explotación de cada uno de ellos y, en consecuencia, puedan saber cuáles son los más adecuados a sus intereses (Rojo, 2010). Alonso (2016), resumiendo las principales características de estos corpus, destaca que la mayoría son escritos, transversales, con textos que desarrollan temas prefijados, especialmente narraciones y textos argumentativos, redactados por informantes habitualmente anglófonos adultos que están aprendiendo español y que tienen distintos niveles de competencia. Pasemos, pues, a relacionarlos y a retratarlos brevemente.

Los corpus orales de ELE que son consultables son:

a. el *Corpus Oral de Español como Lengua Extranjera*[2] (CORELE) (http://cartago.lllf.uam.es/corele/home_es.html.),

b. el *Fono.ELE* (http://www3.uah.es/fonoele/corpus-muestra.php),

c. el *Spanish Corpus Proficiency Level Training* (SPT) (http://www.laits.utexas.edu/spt/intro) y

d. el *Spanish Learner Language Oral Corpora* (SPLLOC1 y SPLLOC2) (http://www.splloc.soton.ac.uk/)[3].

El SPT y el SPLLOC recogen muestra de anglófonos de todos los niveles de dominio de la lengua, mientras que CORELE y *Fono.ELE* contienen muestras de informantes de diferentes lenguas cuyos niveles son de A2 y B1, en el primer caso, y desde A2 hasta C1, en el segundo. Los textos del SPLLOC y de CORELE son entrevistas, descripciones y relatos; los del SPT son diálogos a partir de una serie de preguntas, y *Fono.ELE* contiene grabaciones tanto de breves conversaciones estructuradas como de textos, frases y palabras leídas en voz alta. El SPLLOC y el CORELE cuentan con un corpus de control de hablantes nativos.

Por su parte, los corpus escritos de aprendices de ELE que son consultables son:

a. el *Corpus de aprendices de español* (CAES) (http://galvan.usc.es/caes),

b. el *Corpus del español de los italianos* (CORESPI) (http://corespiycorite.altervista.org/.) y

c. el *Corpus para el análisis de errores de aprendices de ELE* (CORANE) (Cestero y Penadés, 2009)[4].

[2]En su versión inglesa se llama *Spanish Learner Oral Corpus* y puede consultarse en http://cartago.lllf.uam.es/corele/home_en.html.

[3]El CORELE fue realizado por Leonardo Campillos y subvencionado por la Consejería de Madrid y el Fondo Social Europeo. Por su parte, al frente del equipo Fono.ELE están Ana Blanco y Mª Ángeles Álvarez. El SPT fue realizado en el College of Liberal Arts de la Universidad de Texas en Austin bajo el liderazgo de Dale Koike. El *Spanish Learner Language Oral Corpora* fue financiado por el Consejo de Investigación Económica y Social del Reino Unido y llevado a cabo por un equipo integrado por 5 profesoras de las Universidades de York, Newcastle y Southampton.

[4]El CAES es el resultado de un proyecto financiado por el Instituto Cervantes y desarrollado por un equipo de la Universidad de Santiago de Compostela, liderado por Guillermo Rojo e Ignacio Palacios. La autora del CORESPI es Sonia Bailini. Por su parte, el CORANE fue financiado por el Consejo Social de la Universidad de Alcalá y su investigadora principal fue

El CAES y el CORANE contienen textos de aprendices de diversas lenguas maternas (7, en el primer caso, y 23, en el segundo), mientras que el CORESPI es específico de aprendices italohablantes. Los niveles de competencia son distintos en cada corpus: en el CAES encontramos textos desde A1 hasta C1; en el CORESPI, desde A1 hasta B2; y, en el CORANE, desde A2 hasta C1. La tipología textual también varía de un corpus a otro: el CAES es el que presenta mayor diversidad con cartas, postales, correos electrónicos, reservas de hotel, solicitudes de admisión, biografías, narraciones, ensayos y reseñas críticas, todo recogido en línea. En cambio, el CORESPI y el CORANE se circunscriben a un tipo concreto de texto: en el primer caso, interacciones por email entre aprendices y hablantes nativos, y, en el segundo, textos ensayísticos redactados en el aula y fuera del aula. Los tres corpus son longitudinales, pues acogen muestras de los mismos aprendices en varios momentos de su proceso de aprendizaje.

También es accesible el LANGSNAP (http://langsnap.soton.ac.uk/.), que ofrece tanto textos orales como escritos de estudiantes anglófonos. La oralidad se obtiene a través de entrevistas semiestructuradas y narraciones, y las muestras escritas, mediante breves textos argumentativos. Se trata de otro corpus longitudinal, puesto que se recogen 6 textos distintos de cada informante.

Por otro lado, se han confeccionado corpus de aprendices de español, tanto orales como escritos, de los que se conocen algunas de sus características pero que no son de acceso libre, sino que, al menos de momento, son o han sido explotados únicamente por los equipos que los han elaborado. Entre los orales, tenemos el *Corpus Oral de Interlengua Bilingüe Español-Italiano* (CORINÉI), con interacciones nativo/no nativo entre estudiantes de Traducción españoles e italianos a través de Skype, Facebook, Dropbox y similares, tratando diversos temas a lo largo del curso académico (González et al., 2016). Entre los corpus escritos, contamos con el *Anglia Polytechnic University Learner Spanish Corpus* (APU), el corpus *Aprender a Escribir en Lovaina* (Aprescrilov), el *Corpus de Aprendices de Español como Lengua Extranjera* (CAELE), el *Corpus Escrito del Español como L2* (CEDEL2), el *Japanese Learner Corpus of Spanish* (JLCS) y el *Tartu Learner Corpus of Spanish as a L3* (Tartu)[5]. Los corpus que toman muestras

[5]Para más información sobre corpus de aprendices, pueden consultarse el *Learner corpora around the world*, de la Universidad Católica de Lovaina (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html), el *Indexador de Corpus de Aprendices de Español*, desarrollado en la Universidad Complutense de Madrid, (http://repositorios.fdi.ucm.es/corpus_aprendices_espa%C3%B1ol/view/paginas/view_paginas.php?id=1) y el buscador de corpus de la revista electrónica *LinRed* (http://www.linred.es/corpus.html), de la Universidad

de aprendices de una lengua materna específica son el Aprescilov (holandés), el CEDEL2 (inglés), el JLCS (japonés), el Tartu (estonio). En cambio, el APU y el CAELE recogen textos procedentes de estudiantes de español con lenguas maternas diversas. En cuanto a los niveles de competencia, el APU y CEDEL2 contemplan todos los niveles; el CAELE se mueve entre textos de A2 y B1, y el Aprescilov contiene textos desde A1 hasta C1.

3. LOS METADATOS EN LOS CORPUS DE APRENDICES DE ELE

Para entender las posibilidades de explotación que tiene un corpus, además de conocer qué tipos de textos incluye, cómo se han obtenido estos, qué características tienen los aprendices que han proporcionado las muestras orales y/o escritas, es necesario saber cuál es la anotación no lingüística (codificación) que emplea dicho corpus, cuál es su anotación lingüística (etiquetado) y cómo se recuperan los datos.

Los metadatos que tienen en cuenta los corpus de aprendices de ELE que ofrecen aplicativos de consulta automática, metadatos que deben ser codificados y que permiten filtrar la información en las búsquedas automáticas, son de dos tipos: sociolingüísticos y textuales. Los primeros constituyen las habituales variables sociales adscritas o adquiridas que se utilizan en los estudios sociolingüísticos: así, se contempla el sexo o género, la edad y la lengua materna, como variables adscritas, y, como variables adquiridas, podemos encontrar el curso, el centro de estudios, el nivel de dominio de ELE (subjetivo o no), los años de estudio del español, las estancias en países hispanohablantes, el nivel de contacto con la cultura hispana (subjetivo), el nivel de dominio de otras LE (subjetivo o no), la lengua hablada en casa, o las experiencias de aprendizaje. Por su parte, los metadatos textuales incluyen el tipo de tarea que realiza el aprendiz y que genera su muestra de habla oral o escrita, si esta tarea se ejecuta en un contexto controlado o no, o el tipo textual que se requiere al aprendiz.

Pues bien, el corpus *FonoEle* permite extraer datos atendiendo al sexo o género, a la edad, al nivel de dominio de ELE, al nivel de competencia en otra lengua extranjera, y al contacto con el español. El CAES proporciona búsquedas por sexo o género, edad, lengua materna, nivel de competencia de ELE y país de residencia. El CORESPI posibilita filtrar a partir del sexo o género, la edad, el nivel de competencia o dominio de la lengua, otras segundas lenguas conocidas y el tipo textual. El LANGSNAP ofrece los datos

de Alcalá de Henares.

por tarea, por el momento de recogida de la muestra (previa a la estancia en el extranjero, durante la estancia o posteriormente a ella) y por informante; aunque se da información del sexo, la edad, las otras lenguas extranjeras que conoce el aprendiz, los años que lleva estudiando ELE y la condición que haya tenido durante la estadía (estudiante de movilidad, lector, profesor de apoyo), tal cosa se ofrece mediante la descripción que se hace de cada informante, pero no es posible realizar búsquedas automáticas a partir de estas variables sociolingüísticas.

En cuanto a la anotación lingüística, el SPLLOC, el CAES y el CORESPI han sido anotados morfosintácticamente, razón por la cual no solo se pueden buscar palabras, sino también categorías gramaticales y/o estructuras lingüísticas. El CORELE y el CORANE han marcado los tipos de errores, por lo que estos se pueden localizar en el corpus.

4. LOS PROGRAMAS INFORMÁTICOS PARA ANOTAR CORPUS DE APRENDICES DE ELE

En la actualidad, existen diversas aplicaciones informáticas tanto de software libre como comerciales que ayudan en la anotación de corpus lingüísticos. Veamos algunas de ellas y en qué tipo de corpus han sido usadas. La *UAM Corpus Tool* (de 2012, en su versión 2.8) (http://www.corpustool.com/index.html) permite diseñar en árbol las etiquetas específicas que vaya a necesitar el investigador o docente, enlazarlas a los distintos fragmentos que se quieren etiquetar y almacenar el texto etiquetado en XML. Una vez etiquetados los textos, es posible realizar búsquedas aplicando distintos filtros y obtener estadísticas comparativas entre los diferentes subgrupos resultantes de los filtros (frecuencias absolutas y relativas, prueba t de *student* y *chi*^{2}): por ejemplo, saber en qué porcentaje se da el mismo error en cada nivel de dominio de la lengua extranjera. Es una herramienta fácil de utilizar, amigable, de software libre (Figura 1), pero tiene la limitación de que no puede trabajarse de manera holística los errores, sino un tipo concreto de error cada vez (por ejemplo, la confusión entre SER y ESTAR). Ha sido utilizado en estudios de C. Lozano y A. Mendikoetxea[6] a partir del CEDEL2 y con el corpus CAELE (Ferreira y Elejalde, 2017).

[6]Para una relación de las investigaciones de estos autores basadas en la explotación del CEDEL2 y en el uso de *UAM Corpus Tool*, consúltese https://www.uam.es/proyectosinv/woslac/publications.htm.

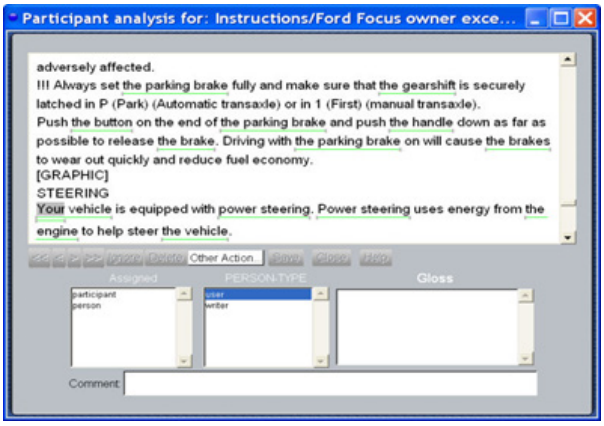

Figura 1. Vista del anotador *UAM Corpus Tool*

El programa *Knowtator* (http://knowtator.sourceforge.net/.) (Figura 2) es una herramienta de anotación de texto de software libre desarrollada en la Universidad de Colorado que utiliza el *backend* basado en archivos Protégé (CLIPS) para guardar la información. Proporciona un mecanismo muy flexible para definir esquemas de anotación, incluida la anotación semántica. Ha sido utilizada por Alonso (2013) aprovechando un conjunto de textos del CEDEL2.

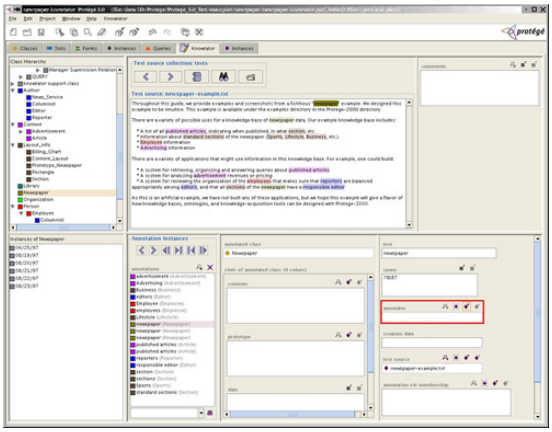

Figura 2. Vista del anotador *Knowtator*

El etiquetador *Markin* (https://www.cict.co.uk/markin/.), también de software libre, facilita la anotación y corrección de textos electrónicos. Puede

importar el texto de un alumno para marcarlo pegándolo desde el portapapeles o directamente desde un archivo de documento. Una vez que se ha importado el texto, *Markin* proporciona un conjunto completo de herramientas que permiten al profesorado marcar y anotar el texto (Figura 3), que queda guardado como un documento XHTML o RTF, que es devuelto al alumno.

Markin tiene un uso muy pedagógico porque, cuando el aprendiz accede al texto anotado, ve claramente los errores, que quedan resaltados en rojo y, situando el cursor en las marcas, puede obtener más detalles sobre la anotación o el comentario del profesor. Esta herramienta ofrece estadísticas del error. Ha sido usado con el corpus *Aprender a Escribir en Lovaina* (Aprescrilov) (Buyse y González, 2013).

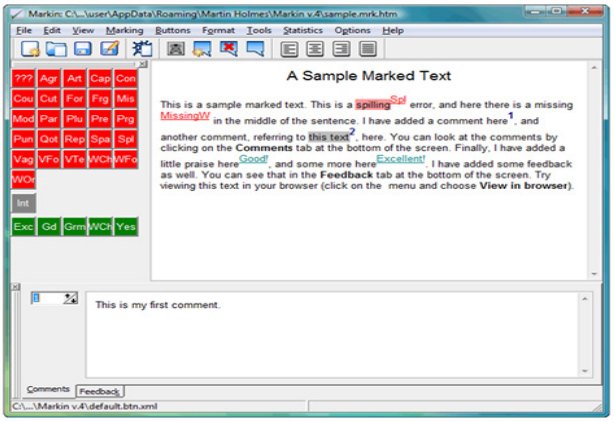

Figura 3. Vista del anotador *Markin*

La interfaz CATMA (https://catma.de/), también de código abierto, ha sido desarrollada por la Universidad de Hamburgo (Figura 4). Marca y analiza textos en web con etiquetas fácilmente definibles. Permite búsquedas interactivas en lenguaje natural, ofrece funciones analíticas estadísticas y no estadísticas automatizadas. Tiene un uso intuitivo y puede utilizarse incluso en idiomas con sistemas de escritura propios (árabe, hebreo, chino). Se utiliza en el CORESPI (Bailini, 2016).

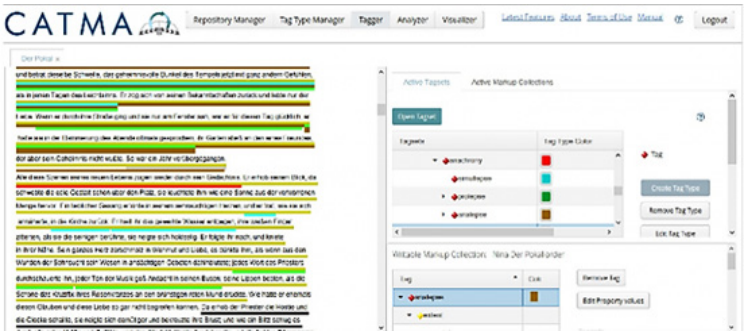

Figura 4. Vista del anotador CATMA

Por último, el programa *Nvivo* (Figura 5) ha sido diseñado por Qualitative Software Research (QSR) para investigaciones basadas en datos cualitativos que procesa de forma cuantitativa y automática (http://www.qsrinternational.com/nvivo-spanish.).

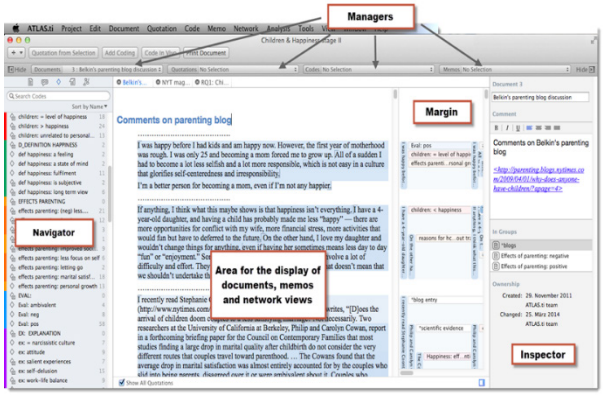

Figura 5. Vista del anotador *Nvivo*

Ayuda en la compilación de textos orales y/o escritos, en su organización y en la creación de la taxonomía de etiquetas para identificar fenómenos. Permite visualizar el contexto de etiquetado, analizar ocurrencias, cruzar variables, aplicar análisis estadísticos. Contiene una opción, 'contexto de etiquetamiento', que muestra el error inserto en el texto. Este anotador lo usan Ferreira, Elejalde y Vine (2014) en un corpus preliminar del CAELE (Ferreira y Elejalde, 2017: 518).

5. EL ETIQUETADOR *TEXTANNOT*

Codificar textos, pero sobre todo etiquetarlos, exige un trabajo colaborativo entre lingüistas e informáticos. El lingüista establece para qué ha de servir el corpus, qué información se quiere obtener, cuáles han de ser los filtros, por tanto, debe tomar decisiones que influirán en la explotación futura que tenga el corpus y en el propio resultado de las investigaciones que se deriven de dicho corpus.

Como hemos visto, la lingüística computacional ha dado ya varios programas informáticos que permiten al investigador etiquetar textos. Existen, incluso, anotadores automáticos, como el *FreeLing*, que, sin embargo, no evitan que haya que aplicar una desambiguación manual. Los etiquetadores existentes están pensados para trabajar un conjunto limitado de textos o una serie limitada de fenómenos. Y los etiquetadores automáticos, basados en la gramática normativa, no acaban de ser adecuados para identificar los errores propios de la interlengua de los aprendices de ELE.

Por estas razones, en el proyecto *Elaboración y catalogación de un corpus de textos escritos en ELE producidos por estudiantes sinohablantes* financiado por el Ministerio de Economía y Competitividad —hoy Ministerio de Ciencia, Innovación y Universidades— (Nº de Ref. FFI2016-80280-R) y desarrollado en la Universidad de Lleida con un equipo de investigadores de esta universidad, de la Universidad de Valencia, de la Universidad de Salamanca, de la Universidad de Estudios Extranjeros de Tianjin, de la Universidad de Changzhou y de la Universidad de Soochow, ha sido necesario diseñar una herramienta específica para poder catalogar y etiquetar de manera semiautomática las muestras escritas y los errores cometidos por aprendices chinos de ELE, el *TextAnnot* (Figura 6). Se trata de una aplicación Web desarrollada con Spring Boot, que es un *framework* o entorno de trabajo proporcionado por API REST (un tipo de interfaz de programación de aplicaciones), que funciona como una arquitectura de *software* que incluye —además de un lenguaje de programación— bibliotecas, programas y otros elementos que sirven para construir y entrelazar los diferentes componentes del proyecto, en este caso, el corpus y el anotador. El código fuente ha sido creado a partir de GitHub, una web que almacena y ayuda a gestionar códigos fuente públicos.

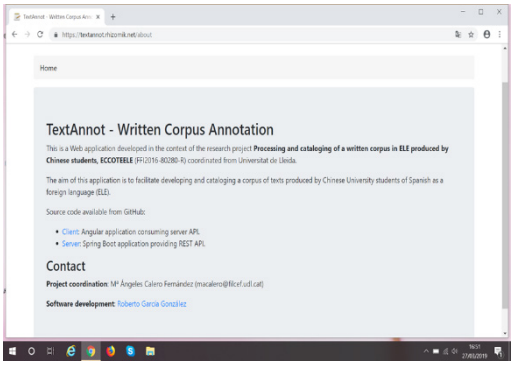

Figura 6. Página de inicio de *TextAnnot*

TextAnnot permite la personalización tanto de los metadatos utilizados para la catalogación de las muestras, como de la jerarquía de etiquetas utilizada para la anotación lingüística. El proyecto es de código abierto y acceso libre (https://github.com/rhizomik/TextAnnot.), tanto para su uso como para colaborar en su desarrollo y mantenimiento. No requiere ningún tipo de instalación: puede accederse a él desde cualquier navegador. Ha sido diseñado por dos miembros del equipo de investigación, que son especialistas en comunicación hombre-máquina. Esta aplicación está en estos momentos en prueba y comenzará a utilizarse en breve para la catalogación, el etiquetado y la consulta del corpus que se está construyendo en el marco del proyecto FFI2016-80280-R, el *Corpus de Interlengua Española de Aprendices Sinohablantes* (CINEAS). En fases posteriores, *TextAnnot* quedará a disposición de los investigadores que necesiten etiquetar los errores de corpus escritos (y orales, si se acompañan de transcripciones) de aprendices, sean de lengua extranjera o de lengua materna.

El CINEAS (https://cineas.udl.cat/.) es un corpus longitudinal constituido por textos redactados por alumnos chinos de Filología Hispánica que cursan sus estudios en las Universidades que participan en el proyecto y que se sitúan entre los niveles A1 y C1, incluidos. Los textos son descriptivos, narrativos, expositivos y argumentativos. Han sido escritos a mano en el aula y han sido recogidos en distintos momentos de cada año académico, desde 2016-2017 hasta 2019-2020. Los metadatos que se recogen son el sexo/género, el curso, la fecha, la universidad en la que se recoge la muestra, el nivel de inglés acreditado, si ha visitado un país hispanohablante y durante cuánto tiempo, así como el género discursivo (Calero, Terrado y Gómez-Devís en prensa). En estos momentos alcanza casi 4000 textos, que están siendo digitalizados en

HTML, catalogados y preparados para su etiquetado en el anotador.

Lo que permite *TextAnnot* es cargar textos individuales o en paquetes, registrar metadatos que faciliten su gestión y procesamiento, señalar en cada texto el fragmento que contiene un error, anotar el tipo de error del que se trata seleccionando del abanico que se ofrece en unos desplegables en forma de árbol —por tanto, en jerarquía—, marcar cada palabra o secuencia incorrecta tantas veces cuantos errores distintos haya en ella, ampliar y modificar la tipología de errores prefijados, revisar las anotaciones, establecer filtros, hacer búsquedas y contabilizar casos (en frecuencias absolutas y relativas). Para establecer un listado preliminar de etiquetas con las que se va a poner a prueba la anotación de errores, se han tenido en cuenta los sistemas de etiquetas utilizados en diferentes corpus, así como los establecidos en estudios sobre el error en aprendices de lenguas extranjeras, algo que no podemos tratar aquí por falta de espacio.

6. CONCLUSIONES

Tras esta revisión de los corpus de aprendices de ELE, puede concluirse que la destreza que mejor puede caracterizarse es la expresión escrita, al menos así lo entendemos dado que la mayoría de los corpus son escritos. Los metadatos que se utilizan y que, en algunos casos, sirven para filtrar la información susceptible de extraerse del corpus, son especialmente el sexo o género, la edad, la lengua materna (cuando son corpus multilingües), el nivel de dominio del español o el conocimiento de otra(s) lengua(s) extranjera(s). Este tipo de metadatos hace posible analizar los elementos lingüísticos distinguiendo a los informantes por rasgos sociales, lo que facilitaría la detección de eventuales influencias de estos factores en el aprendizaje de ELE, como sería previsible si atendemos a los resultados de los estudios sociolingüísticos tanto de comunidades monolingües como bilingües. Por otro lado, la anotación morfosintáctica que han recibido el SPLLOC, el CAES y el CORESPI y el etiquetado de tipos de errores del CORANE y el CORELE, pueden ayudar a identificar fenómenos de la interlengua.

El CEDEL2, el CAELE, el CORESPI y el Aprescrilov han utilizado programas de software libre distintos, más o menos intuitivos y flexibles, que sirven para diferentes tipos de etiquetado (de categorías gramaticales, de errores, semántico…) y que proporcionan cálculos estadísticos de aquello que se busca. El *TextAnnot* está pensado para el análisis de errores, y aporta una

tipología de errores prefijada (no hay que crearla) que puede ser aprovechada en su diseño original o puede ser adaptada a las necesidades de investigación o docentes y a las características de las muestras que forman el corpus en el que se aplique. Se está empleando por primera vez en el CINEAS, pero podrá utilizarse en cualquier corpus de aprendices, tanto de lengua extranjera como de lengua materna. Un anotador de estas características posibilitaría la comparación de datos entre diferentes corpus y aumentaría el impacto de los resultados obtenidos en la planificación didáctica de ELE, en general o destinada a colectivos específicos de aprendices.

REFERENCIAS BIBLIOGRÁFICAS

Ainciburu, Mª. C. 2012. "Recursos para estudiar el español a partir de corpus", *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, 12: 209-215.

Alonso Ramos, M. 2013. "Colocaciones, diccionarios y corpus de aprendices", *Eugenio Coseriu, in mermoriam, XIV Jornadas de Lingüística*, M. Casas Gómez (dir) y R. Vela Sánchez (ed.). Cádiz: Universidad de Cádiz. 57-71.

Alonso Ramos, M. 2016. "Spanish learners corpus research. Achievements and challenges", *Spanish Learner Corpus Research. Current trends and future perspective*s, M. Alonso-Ramos (ed.). Amsteerdam/Philadelphia: John Benjamins. 3-32.

Bailini, S. 2016. *La interlengua de lenguas afines. El español de los italianos, el italiano de los españoles*, Milano: Edizioni Universitarie di Lettere Economia Diritto.

Baralo Ottonello, M. 2010. "La investigación en español como lengua segunda: necesidad de un corpus de español no nativo". *V Congreso Internacional de la Lengua Española (CILE)*, Valparaíso (Chile). Recuperado a partir de: http://congresosdelalengua.es/valparaiso/ponencias/lengua_educacion/baralo_marta.htm

Buyse, K. 2011. "¿Qué textos en línea utilizar para qué fines en el aula de ELE?", *Del texto a la lengua: La aplicación de los textos a la enseñanza-aprendizaje del español L2-LE,* vol I, J. de Santiago Guervós, H. Bongaerts, J. J. Sánchez Iglesias, M. Seseña (coords). Madrid: Arco-Libros. 277-288.

Buyse, K. y González Melón, E. 2013. "El corpus de aprendices Aprescrilov y su utilidad para la didáctica de ELE en la Bélgica multilingüe", *Plurilingüismo y enseñanza de ELE en contextos multiculturales. XXIII Congreso Internacional ASELE*, en B. Blecua, S. Borrell, B. Crous y F. Sierra (eds.). Girona: ASELE y Universidad de Girona. 247-261.

Buyse, K., Fernández Pereda, L., Verveckken y K. 2016. "The Aprescrilov corpus, or broadening the horizon of Spanish language learning in Flanders", *Spanish Learner Corpus Research. Current trends and future perspective*s, M. Alonso-Ramos (ed.). Amsteerdam/Philadelphia: John Benjamins. 143-168.

Calero Fernández, Mª. Á., Terrado Pablo, F. J. y Gómez-Devís, M-B.

en prensa. "Análisis de errores y variación: la importancia de los corpus lingüísticos", *Variación lingüística y su didáctica. Actas del Congreso Internacional sobre Variación Lingüística en Español: Cómo trabajarla en el aula* (Salamanca, 28 y 29 de junio de 2017). Universidad de Salamanca.

Cestero Mancera, A. Mª., Penadés Martínez, I. 2009. *Corpus de textos escritos para el análisis de errores de aprendices de E/LE (CORANE)*. Alcalá de Henares: Universidad de Alcalá de Henares. CD-ROM.

EAGLES. 1996. *Text Corpora Working Group Reading Guide. Eagles* (Expert Advisory Group on Language Engineering). Recuperado de http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html

Ferreira Cabrera, A., Elejalde Gómez, J., Vine Jara y A. 2014. "Análisis de Errores Asistido por Computador basado en un Corpus de Aprendientes de Español como Lengua Extranjera", *Revista Signos. Estudios de Lingüística*, 47(86): 385-411.

Ferreira Cabrera, A. y Elejalde Gómez, J. 2017. "Análisis de errores recurrentes en el Corpus de Aprendices de Español como Lengua Extranjera, CAELE", *Revista Brasileira de Lingüística Aplicada*, 17: 509-538.

González Royo, C., Chiapello, S., Martín Sánchez, T., Mura, G. Á., Pascual Escagedo, C., Puigdevall Bafaluy, N. y Regagliolo, A. 2016. "CORINÉI (Corpus Oral de Interlengua Bilingüe Español-Italiano): Elaboración, análisis y aplicación a la enseñanza/aprendizaje de la interacción con las TICs", *Innovaciones metodológicas en docencia universitaria: resultados de investigación*, por J. D. Álvarez Teruel, S. Grau Company, Mª. T. Tortosa Ybáñez (coords.). Universidad de Alicante. 1951-1958.

Leech, G. 1993. "Corpus Annotation Schemes", *Literary and Linguistic Computing*, 8(4): 275-281.

Penadés Martínez, I. 2005. "Corpus para el análisis de errores en el aprendizaje de ELE. Presentación", *Linred*, 3: 1-5.

Pérez Ávila, E. 2007. "El corpus lingüístico en la didáctica del léxico del español como LE", *Boletín de la Asociación para la Enseñanza del Español como Lengua Extranjera,* 37: 11-27.

Rojo, G. 2010. "Sobre codificación y explotación de corpus textuales: otra comparación del Corpus del Español con el CORDE y el CREA",

Lingüística, 24: 11-50.Ardila A. 2011. “Neuropsicología del lenguaje”, *Manual de Neuropsicología*, J. Tirapu Ustárroz, M. Rios Lago, F. Maestú Unturbe (eds.). Barcelona: Viguera. 99-121.