MENDIVE
REVISTA DE EDUCACIÓN

Artículo de revisión

Diseño teórico y metodológico del Corpus de Aprendientes Chinos de Español

Theoretical and methodological design of the Corpora of Chinese Learners of Spanish

Desenho teórico e metodológico do Corpus de Chineses Aprendizes de Espanhol

Jialing Gou[1]
https://orcid.org/0000-0002-0303-974X
Diusbel Rodríguez Roque[2]
https://orcid.org/0009-0003-4224-8928
Lidia Ester Cuba Vega[1]
https://orcid.org/0000-0002-0405-5194

[1] Universidad de La Habana. Cuba.
wawaana0601@gmail.com, lidiacuba@fenhi.uh.cu
[2] Sichuan International Studies University. China
diusbelrr2010@gmail.com

Recibido: 27 de enero 2023
Aceptado: 25 de mayo 2023.

RESUMEN
Este trabajo versa sobre los corpus de aprendientes y forma parte de una investigación doctoral titulada *La atenuación en la expresión e interacción escrita de estudiantes chinos de español,* cuyo objetivo (caracterizar el empleo de la atenuación en la producción escrita de estos estudiantes) y la carencia de datos que permita cumplirlo, conduce al diseño un corpus de aprendientes, fuente de datos empíricos fiable para el análisis de las características de la interlengua de estos estudiantes. Partiendo de la idea de que un corpus siempre responde a un diseño previo, este trabajo persiguió como objetivo diseñar teórica y metodológicamente un corpus informatizado, representativo de las variantes oral y escrita del español como lengua extranjera y segunda de los estudiantes chinos, para contar con una base de datos que permita efectuar investigaciones rigurosas sobre su interlengua e identificar las principales dificultades en su aprendizaje, a fin de diseñar estrategias didácticas adecuadas para solucionarlas. Se asumió el Método Dialéctico-Materialista, como método general de las ciencias y sustento filosófico, y se emplearon métodos teóricos y empíricos, a saber, el histórico-lógico, el análisis-síntesis, el sistémico-estructural, el análisis documental y la modelación. Los resultados principales exhiben la sistematización teórica sobre lingüística del corpus y corpus de aprendientes y, a partir de ella, el diseño teórico y metodológico de un corpus especializado, muestral y representativo, multilingüe, mixto, abierto y de orientación transversal; además, la concepción de tres etapas para la creación de Corpus de Aprendientes de Chinos de Español: diseño, recolección de datos y construcción.

Palabras clave: lingüística de corpus; corpus de aprendientes; estudiantes chinos; español como lengua extranjera/segunda lengua.

ABSTRACT
This paper deals with corpora of learners and is part of a doctoral research entitled Attenuation in the written expression and interaction of Chinese learners of Spanish, whose objective (to characterize the use of attenuation in the written production of these learners) and the lack of data to achieve it, leads to the design of a corpora of learners, a source of reliable empirical data for the analysis of the characteristics of the interlanguage of these learners. Starting from the idea that a corpus always responds to a previous design, this work pursues the theoretical and methodological design of a computerized corpora, representative of the oral and written variants of Spanish as a foreign and second language of Chinese learners, in order to have a database that allows us to carry out rigorous research on their interlanguage and identify the main difficulties in their learning, in order to design appropriate didactic strategies to solve them. The Dialectical-Materialist Method is assumed as the general method of sciences and philosophical support, and theoretical and empirical methods are used, namely, historical-logical, analysis-synthesis, systemic-structural, documentary analysis and modeling. The main results show the theoretical systematization on linguistics of corpora and corpora of learners and, from it, the theoretical and methodological design of a specialized, sample and representative, multilingual, mixed, open and transversal-oriented corpus; in addition, the conception of three stages for the creation of Corpora of Chinese Learners of Spanish: design, data collection and construction.

Keywords: linguistics of corpora; corpus of learners; Chinese learners; Spanish as a foreign/second language.

RESUMO
Este trabalho trata do corpus de aprendizes e faz parte de uma pesquisa de doutorado intitulada Atenuação na expressão e interação escrita de estudantes chineses de espanhol, cujo objetivo (caracterizar o uso da atenuação na produção escrita desses estudantes) e a falta de de dados que permite a conformidade, leva ao desenho de um corpus de alunos, uma fonte de dados empíricos confiável para a análise das características interlínguas desses alunos. Partindo da ideia de que um corpus responde sempre a um desenho prévio, este trabalho teve como objectivo conceber teórica e metodologicamente um corpus informatizado, representativo das variantes orais e escritas do espanhol como língua estrangeira e segunda de estudantes chineses, para ter uma base de dados que permite pesquisar rigorosamente a sua interlíngua e identificar as principais dificuldades na sua aprendizagem, a fim de desenhar estratégias didáticas adequadas para resolvê-las. Assumiu-se o Método Dialético-Materialista, como método geral de ciência e sustentação filosófica, e foram utilizados métodos teóricos e empíricos, nomeadamente, histórico-lógico, análise-síntese, sistêmico-estrutural, análise documental e modelagem. Os principais resultados expõem a sistematização teórica sobre linguística do corpus e corpus de aprendizes e, a partir dela, o desenho teórico e metodológico de um corpus especializado, amostral e representativo, multilíngue, misto, aberto e de orientação cruzada; Além disso, a concepção de três etapas para a criação de Corpus de Aprendizes de Chinês de Espanhol: concepção, coleta de dados e construção.

Palavras-chave: linguística de corpus; corpus de aprendizes; estudantes chineses; Espanhol como língua estrangeira/segunda língua.

INTRODUCCIÓN

En el área de la lingüística general y la aplicada en particular, se procura trabajar, hoy en día, con datos reales y significativos que posibilitan un análisis lo más fiel posible de su objeto de estudio. Esto conlleva a la recopilación, organización y codificación adecuada de muestras extensas de la realidad lingüística que se pretende investigar. En este sentido, las Tecnologías de la Información y las Comunicaciones (TICs) ha facilitado, de manera considerable, la construcción de corpus en la medida en que han hecho posible la recopilación y organización de textos en formato electrónico, lo que permite además su análisis, codificación y anotación a través de aplicaciones desarrolladas para estos fines y, por consiguiente, favorece la recuperación de información estadística y muestras de lengua mediante una herramienta informática de búsqueda. Como apunta Rojo (2022), «la difusión de las computadoras, así como la generalización del uso de Internet, han supuesto una auténtica revolución instrumental en lingüística y la lingüística de corpus es, probablemente, su exponente más claro» (p. 73).

Los corpus se caracterizan por estar compuestos por una voluminosa cantidad de datos reales representativos de una lengua, variedad lingüística, dialecto, etc. o de la interlengua de estudiantes de una determinada lengua extranjera (LE) o segunda lengua (L2), que se almacenan en formato electrónico y se codifican para poderlos analizar científicamente, de modo que se pueda arribar a resultados empíricos que permitan estudiar en profundidad fenómenos y elementos lingüísticos a fin de comprender mejor el lenguaje y las lenguas, y establecer relaciones entre teorías y realidades lingüísticas. Pero, para construir un corpus que facilite la recuperación de datos homogéneos y cuantificables que favorezcan la elaboración de teorías empíricas acerca de un objeto de estudio determinado, resulta imprescindible un diseño adecuado.

En este artículo se parte de la afirmación de Rojo (2021) de que «todos los aspectos necesarios para comprender qué es un corpus […] se sintetizan en la idea de que siempre responden a un diseño previo» (p. 81). Y, precisamente, en estas páginas se efectúa una sistematización teórica sobre lo que se entiende hoy por lingüística de corpus, corpus lingüístico y corpus de aprendientes, así como una revisión de los principios metodológicos que rigen la constitución de un corpus injustico para fundamentar desde el punto de vista teórico y metodológico la construcción del Corpus de Aprendientes Chinos de Español (CACHE). Pero ¿por qué es necesaria la creación de este corpus?

En los últimos años se ha construido un número considerable de corpus de aprendientes de ELE/L2. Sin embargo, todavía son pocos los que incluyen muestras de interlengua española de informantes chinos. Esto se puede constatar en los repositorios de corpus de aprendientes de español de la Universidad Católica de Lovaina y de la Universidad Complutense de Madrid (UCM). Una revisión de estos repositorios permite constatar que los corpus que incluyen muestras de la interlengua de aprendientes chinos de español son el Corpus Oral de Español como Lengua Extranjera (CORELE), FonoELE, DIAZ Corpus, Corpus de Interlengua Española de Aprendices Sinohablantes (SINEAS), el Corpus Escrito del Español como L2 (CEDEL2), el 台灣西語學習者語料庫 (Corpus de Aprendices Taiwaneses de Español, CATE) y el Corpus Aprendices de Español (CAES)

De ellos, los tres primeros (CORELE, FonoELE y DÍAZ Corpus) constituyen corpus orales monolingües que recogen muestras de la interlengua producida por estudiantes de español con diferentes lenguas maternas. Cabe señalar que los criterios del diseño de cada uno responden a los objetivos con que fueron creados y, por lo tanto, son distintos. Por ejemplo, CORELE fue recogido para el estudio de análisis de errores de la producción oral y FonoELE para el desarrollo de investigaciones sobre la adquisición y aprendizaje del componente fónico en ELE.

Todos recogen muestras de la interlengua de hablantes de diferentes lenguas maternas, pero las muestras de informantes chinos son minoritarias frente a las de hablantes de lenguas europeas como el portugués, italiano y el alemán.

Esto último no afectaría la realización de investigaciones con las muestras de hablantes chinos si hubiera una representación de todos los niveles de competencia y de poderse establecer una muestra homogénea en este sentido, pero, CORELE solo está compuesto por discursos de hablantes acreditados con los niveles de competencia A2 y B1, FonoELE incluye desde el nivel A2 hasta C1 y Diaz Corpus no declara este metadato. Además, como los criterios del diseño y los objetivos de estos corpus son distintos y únicamente FonoELE documenta y describe elementos del diseño del corpus ofreciendo información relevante sobre su construcción, resulta imposible realizar estudios contrastivos con estas bases de datos empíricos informatizada.

De los restantes corpus listados, el SINEAS y CAES constituyen corpus escritos. El SINEAS incluye solamente muestras de la interlengua española de informantes chinos, mientras que el CAES incorpora la de estudiantes de variadas lenguas primeras. Por su parte, 台灣西語學習者語料庫(CATE) solamente incluye la lengua producida por estudiantes de español de Taiwán y CEDEL2 incluye estudiantes de variadas lenguas extranjeras, dentro de las que se incluye el chino mandarín. Cabe apuntar que estos dos últimos corpus son mixtos, es decir, contienen textos orales y escritos; además, son multilingües pues contienen corpus de nativos de las lenguas primeras de los estudiantes aprendientes de español, una característica de gran valor para el desarrollo de estudios contrastivos.

El Corpus de Interlengua Española de Aprendices Sinohablantes (SINEAS) recoge textos escritos por estudiantes chinos de la carrera de Filología Hispánica de cuatro universidades chinas y de estudiantes de nacionalidad china en situación de movilidad (en inmersión lingüística) en tres universidades españolas desde el curso 2016-2017 hasta el curso 2019-2020. Las muestras son de estudiantes con diferentes niveles de dominio lingüístico, de A1 a C1. Los textos recogidos, en tipos textuales diferentes (descripción, narración, exposición y argumentación), fueron escritos a mano a partir de tareas de escritura como estímulo integradas en 43 pruebas, luego fueron digitalizados respetando fielmente la versión manuscrita, y han sido anotados a partir de un conjunto de metadatos sociales y lingüísticos que permitan filtrar las búsquedas. Estadísticamente, se puede apuntar que CINEAS (la versión 1.0, diciembre de 2019) contiene muestras producidas por estudiantes chinos con un total de 4378 textos escritos 4339 (su volumen actual es de 435.000 palabras). Se puede acceder solo a una parte del corpus CINEAS, el acceso a la totalidad está restringido. Solamente podrán acceder a todos los datos dispuestos en CINEAS los investigadores e investigadoras del proyecto de la Universidad de Lleida. El público general solamente puede hacer consultas simples, sin el etiquetado de errores y otros metadatos, de menos del 50% del corpus.

El Corpus Aprendices de Español (CAES), en su versión actual (2.1, marzo 2022), recoge 1 045 097 elementos lingüísticos, producidos por 2544 estudiantes de diferentes niveles de lengua (A1-C1) y diferentes L1 (alemán, árabe, chino mandarín, francés, griego, inglés, italiano, japonés, polaco, portugués y ruso), que escribieron de 2 a 3 textos cada uno, lo que comprende unas 6561 tareas de redacción integradas en 2544 pruebas. Este gran corpus, con acceso abierto, ha sido construido gracias a la recogida de muestras en distintos centros del Instituto Cervantes y universidades de un gran número de países, en un período que va desde octubre de 2011 que inició este proyecto promovido y financiado por el Instituto Cervantes hasta diciembre de 2020.

Aunque, sin lugar a duda, el CAES constituye una gran base de datos de la interlengua española producida por estudiantes chinos, toda vez que

actualmente incluye 108 942 unidades lingüísticas distribuidas por diferentes niveles (A1, 19 751; A2, 23 292; B1 48 766; B2, 16 892; y C1, 241) producida por este estudiantado de ELE/El2 y está constituido por una gran variedad de textos (biografías, cartas, correos electrónicos, ensayos argumentativos, narraciones, notas/avisos, postales, etc.), hay que señalar al menos tres aspectos que se consideran negativos a la hora de realizar estudios con este corpus.

Primeramente, las tareas de escritura a partir de las que se recogieron las muestras de interlengua de este corpus imitan, en gran medida, las pruebas de DELE, que, a juicio de esta autora, dada su experiencia en la preparación de estudiantes para realizar los exámenes de DELE, y en la revisión de temarios de los Cursos de Acreditación de Examinadores de DELE realizados, limitan la creatividad de los estudiantes y proveen solamente una interlengua condicionada por unas pautas preestablecidas y de obligatorio cumplimiento. Por otra parte, como se recogen muestras en variados países, habría estudiantes de ELE y EL2, así como estudiantes de español como lengua de herencia, etc. Así pues, la no indicación del metadato referido al contexto de inmersión o no inmersión donde aprenden estos estudiantes, o el tiempo de contacto con la lengua española como años de estudio o estancias en el extranjero, limita el desarrollo de estudios contrastivos y sociolingüísticos a partir de estas variables. Por otra, se desconoce el nivel de instrucción de los estudiantes y los programas en los que aprenden español, sería inapropiado comparar la interlengua de español de un estudiante chino de la carrera en Filología Hispánica y la de un estudiante de un curso de español de 6 meses, aunque ambos estén acreditados con un nivel A1.

El 台灣西語學習者語料庫 (Corpus de Aprendices Taiwaneses de Español, CATE), como muestran sus dos subcorpus constituyentes, 台灣西語學習者書面語語料庫 (Corpus Escrito de Aprendices Taiwaneses de Español, CEATE) y el 台灣西語學習者口語語料庫 (Corpus Oral de Aprendices Taiwaneses de Español, COATE), es un corpus oral y escrito que recoge la lengua española producida por estudiantes de Taiwán. Constituye una base de datos valiosa con unas 34 000 palabras producidas por estudiantes de ELE de Taiwán, sin embargo, su acceso está restringido. Además, pueden señalarse dos aspectos negativos de cara a la investigación. Primeramente, este corpus no indica el metadato referido a la lengua de origen y, aunque cabe suponer que la mayoría de los textos proceden de informantes cuya L1 es el chino mandarín, si se considera la situación sociolingüística de Taiwán, la L1 también podrían ser otras como el taiwanés (*min nan*) y hakka (*kejia*). Además, puede señalarse que la representatividad de la interlengua de estos estudiantes se ve afectada, toda vez que solo se incluyen textos ensayísticos y, como consecuencia, no contine la diversidad de tipos de textos que exige un corpus de aprendientes.

El Corpus Escrito del Español como L2 (CEDEL2) es un gran corpus lingüístico que contiene textos producidos por estudiantes de español como segunda lengua / lengua extranjera. Su nueva versión (CEDEL2, versión 2, 2021) cuenta actualmente con un total de 4.399 participantes y 1.105.936 palabras (744,950 de estudiantes de ELE/L2 y 360,986 de nativos), lo que lo convierte en uno de los mayores corpus de su tipo. Este corpus de aprendientes contiene datos de estudiantes de español con diferentes L1 (lengua materna de los estudiantes): inglés, alemán, holandés, francés, portugués, italiano, griego, ruso, japonés, árabe y el chino. También, para fines comparativos, CEDEL2 contiene subcorpus de textos producidos por hablantes nativos del español (variedades españolas e hispanoamericanas), inglés, japonés, portugués, griego y el árabe. Todavía no contiene muestras de informantes nativos del chino mandarín.

Pese a que CEDEL2, a diferencia de CINEAS y CATE, y siguiendo la filosofía *open data science,* está disponible de forma íntegra y se puede consultar en línea y descargar libremente para fines investigativos y de enseñanza, todavía no recoge una muestra

extensa y representativa de textos producidos por estudiantes de español cuya lengua materna sea el chino mandarían. Al observar las estadísticas que se ofrecen en su sitio web oficial, se constata que, de los 4399 informantes, solamente hay 22 chinos, 20 de ellos proporcionaron textos escritos y 2 participaron en la recogida de textos orales. Estas cifras, resultan insignificante al lado de los datos recogidos con informantes de otras L1, por ejemplo, con inglés como L1 participaron 1931, con japonés 243, con griego 216, con portugués 164 y con ruso 101. Además, no se ha conformado el corpus de nativos con datos lingüísticos del chino mandarín.

Aún más, los datos de la interlengua de estudiantes chinos recopilados en CEDEL2 resultan insuficientes para la realización de estudios rigurosos. Por una parte, no es una muestra representativa de la interlengua de estos estudiantes, ya que no hay informantes de todos los niveles de competencia y de los niveles representados el número de informantes es ínfimo. Por otra, también debe señalarse que de los 20 textos escritos recopilados, 17 pertenecen a informantes mujeres y solo 3 a hombres. Por tanto, es imposible conformar una muestra homogénea que permita efectuar un estudio teniendo en cuenta factores sociales y lingüísticos fundamentales para los propósitos de esta investigación. También puede indicarse que los 20 textos de informantes chinos fueron escritos a partir de una misma tarea, que exigía la redacción de una descripción-narración, por lo que tampoco hay representatividad de géneros textuales.

Hasta ahora no se dispone de un corpus, entre los corpus de aprendientes de ELE existentes, que sirva para investigar profundamente y con rigor la interlengua española de los estudiantes chinos, teniendo en cuenta ciertas variables lingüísticas y extralingüísticas. El CACHE viene a cubrir ese vacío.

La fuerza que está tomando el aprendizaje del español como lengua extranjera en China, la poca profundización de algunas investigaciones sobre la interlengua que desarrollan los estudiantes chinos -debido en gran medida a que los datos empíricos utilizados para el análisis son reducidos y los instrumentos de recogida de datos son inadecuados- y, sobre todo, la inexistencia de un corpus para el estudio de la interlengua de estudiantes de español que ofrezca una muestra extensa de textos orales y escritos producidos por estudiantes de origen chino hablantes de mandarín y con etiquetas de metadatos que indiquen información precisa de los factores sociales y lingüísticos de los informantes que permita extraer conclusiones realmente generalizables de manera rigurosa, justifican el diseño y construcción del CACHE, que, además, deberá ser de acceso abierto.

Objetivo general: Diseñar teórica y metodológicamente un corpus informatizado, representativo de las variantes oral y escrita del español como lengua extranjera y segunda de los aprendientes chinos, para contar con una base de datos que permita llevar a cabo investigaciones rigurosas sobre su interlengua e identificar las principales dificultades en su aprendizaje, a fin de diseñar estrategias lingüodidácticas adecuadas para su solución.

DESARROLLO

Lingüística de corpus, ¿una nueva teoría, una (sub)disciplina lingüística o una nueva metodología?

«La lingüística es una ciencia empírica cultural […] cuyo objeto de estudio son el lenguaje y las lenguas» (Rojo, 2021, p. 35). Pese a que esta consideración es reciente, hoy en día es ampliamente aceptada. Sin embargo, la profundización teórica sobre la lingüística de corpus (LC), una realidad joven (seis décadas de historia), deja a la vista cuestiones terminológicas sobre las que coexisten de forma explícita o implícita ambigüedades conceptuales, vaguedades terminológicas irresueltas y, por consiguiente, la

convergencia de múltiples perspectivas opuestas.

Precisamente, el primer problema terminológico radica en definir qué es la LC. ¿Estamos ante una nueva teoría?, ¿constituye una (sub)disciplina o se trata, más bien, de una nueva metodología? Se ha debatido mucho para dar respuesta a estas preguntas; sin embargo, todavía no se observa mucho consenso en la bibliografía. Una muestra de ello se constata en la divergencia de criterios entre los autores que reflexionan sobre las cuestiones teóricas de la LC. La respuesta más acertada parece ser la ofrecida por Rojo (2021), quien afirma que:

> En efecto, la LC no es una teoría: los datos procedentes de un corpus pueden ser analizados desde muy diferentes marcos teóricos […]. Tampoco parece que pueda ser considerada como una (sub)disciplina lingüística como la morfología, la sintaxis, la sociolingüística, etc.: la utilización de corpus se da en estudios gramaticales, históricos, sociolingüísticos, lexicográficos y muchos otros campos especializados. Por último, no es sencillo contemplarla como una metodología en el sentido más estricto de la palabra. (Rojo, 2021, p.47)

El propio Rojo (2021) refiriéndose a la dificultad de contemplar a la LC como una metodología señala que «Leech (1992, p. 106) considera que la LC es "a new research enterprise, and in fact a new philosophical approach to the subject"» (p. 47) y más adelante enfatiza en que «más recientemente, Leech (2011, p. 158) ha insistido en considerar que la LC "is not a methodology pure and simple, but is more like a methodology than a scientific domain"» (p.47). Se observa, pues, la complejidad que han advertido autores como Leech a la hora de definir qué es la LC y la ambigüedad de sus propuestas.

De acuerdo con Rojo (2021) una idea clara y operativa del carácter de la LC es considerar que se trata de una aproximación al estudio de los hechos lingüísticos de orientación empírica y basada en el análisis detallado de gran cantidad de datos (los corpus), con lo que queda patente su oposición tanto a la lingüística racionalista como a la descriptiva tradicional […]. En efecto, la LC constituye una forma de acercamiento al estudio de los fenómenos y elementos lingüísticos fundamentada en ciertos supuestos acerca de qué aspectos del análisis son realmente relevantes. (p. 48)

Rojo (2021), tomando como referencia las propuestas de Leech (1992), Biber, Conrad y Reppen (1998), Tognini-Bonelli (2001), Gries (2006, 2009) y Bolaños (2015), considera que la LC se caracteriza fundamentalmente por los siguientes rasgos:

- Ser empírica, centrarse en el análisis de los esquemas reales de uso en textos producidos de forma natural. Por tanto, la LC está más interesada en la actuación que en la competencia y en la descripción de lo que se encuentra en las lenguas que en los universales lingüísticos.
- Utilizar corpus textuales amplios como base del análisis. En la medida en que esos corpus estén bien construidos y sean representativos, lo que se encuentra en la muestra puede ser proyectado a la población, es decir, a la lengua. El objetivo, por tanto, no es solo describir y explicar lo que se encuentra en el corpus, sino todo lo que se puede ver en la lengua o variedad lingüística de la que ha sido extraído.
- Los datos proceden de textos producidos de forma natural, lo cual permite poner en relación los datos obtenidos con las variaciones debidas a los diferentes registros y tipos de texto, relación que no resulta posible en los datos

obtenidos mediante los diseños experimentales. Naturalmente, las variaciones pueden estar situadas en las líneas diacrónica, diatópica y diastrática.

- Hacer un uso intenso de computadoras para llevar a cabo al menos una parte de ese análisis. La utilización de recursos computacionales en la construcción y explotación de los corpus es una necesidad derivada de su volumen, puesto que solo así es posible analizar conjuntos de datos de un tamaño que resultaría imposible sin esos recursos.
- Emplear técnicas analíticas de carácter cuantitativo y cualitativo. Es importante señalar que los corpus textuales son el recurso más cómodo y adecuado para estudiar todo lo relacionado con la frecuencia de los fenómenos y elementos lingüísticos.
- Realizar (o, al menos, pretender realizar) análisis sistemáticos y exhaustivos de todos los casos pertinentes localizados en el corpus de aquello que se pretende estudiar. Es decir, el corpus no es tratado simplemente como una especie de base de datos de la cual se extraen unos cuantos casos y se rechazan otros, sino que se toma en consideración la totalidad del corpus.

Por lo tanto, resumiendo el posicionamiento de Rojo (2021) se puede concluir que la LC es una aproximación empírica que analiza datos reales - «what people actually say and write» (Aarts, 2002, p. 4)-, representativos de los actos lingüísticos ocurridos en una comunidad lingüística, con la idea de comprender el sistema que los ha hecho posibles. Y, el modo de lograrlo es examinando los textos o fragmentos de textos, orales o escritos, contenidos en corpus representativos, cuya exhaustividad obliga a su construcción informatizada, a trabajar con computadoras para la extracción y recuperación selectiva de la información y poder, finalmente, realizar el procesamiento estadístico de esas grandes masas de información.

Igualmente, Rojo (2021) considera que la metodología de la LC también cumple con las características del "ciclo empírico" de Krug, Schlüter y Rosenbach (2013):

> En primer lugar, la objetividad, lo cual significa que los datos utilizados tienen que ser por completo independientes de las personas que realizan la investigación y de las herramientas utilizadas en su obtención. En segundo término, la fiabilidad y replicabilidad, que garantizan que los datos obtenidos serán los mismos en extracciones realizadas en momentos diferentes. Por último, la relevancia de los datos utilizados para el fenómeno analizado. (2021, p. 50)

De modo que, hay que considerar que la LC es una orientación empírica que se ocupa del análisis de los datos reales y objetivos reunidos en corpus lingüísticos, valiéndose de nuevas tecnologías y programas informáticos. Constituye, así, una orientación empírica en la que «el crecimiento cuantitativo de los conocimientos sobre el comportamiento de las lenguas y los hablantes ha dado lugar a un importante crecimiento cualitativo» (Rojo, 2021, p. 50).

¿Qué es un corpus lingüístico?

En general, asumir una definición operacional como marco teórico de una investigación resulta una tarea compleja. Y, por supuesto, la heterogeneidad de acercamientos a la hora de definir qué se entiende como corpus dentro de la LC es una muestra de dicha complejidad. Una revisión en profundidad de la bibliografía especializada viabiliza la presente sistematización crítica con la intención de asumir aquella que exponga las

características más relevantes de un corpus dentro de la LC actual o para poder construir una definición rigurosa para el presente estudio.

Tras consultar las definiciones de corpus lingüísticos ofrecidas por Francis, Kuèera y Mackie (1982), Sinclair (1991, 2005), Leech (1992), Biber (1993), Torruella y Llisterri (1999), McEnery, Xiao y Tono (2006), Parodi (2008), Villayandre (2008), Hincapié (2018), Hincapié y Bernal (2018), Hincapié y Rubio (2018), Lemnitzer y Zinsmeister (2008) y Rojo (2021), se considera que la definición ofrecida por este último autor incluye, de forma resumida, los rasgos más característicos de un corpus lingüístico en la LC actual. Según Rojo (2021), un corpus es:

> … un conjunto de (fragmentos de) textos, orales o escritos, producidos en condiciones naturales, conjuntamente representativos de una lengua o una variedad lingüística, en su totalidad o en alguno(s) de sus componentes, que se almacenan en formato electrónico y se codifican con la intención de que puedan ser analizados científicamente. (p. 1)

El propio Rojo (2021) reflexiona sobre los aspectos que recoge en esta definición. A continuación, se comenta su análisis.

Según este lingüista, «los textos que integran el corpus deben haber sido producidos en situaciones naturales» (2021, p. 1). Seguidamente explica que, antes de incluir los textos o fragmentos de textos en el corpus, deben haber sido creados como una obra literaria, un texto periodístico, un texto epistolar (en el caso de los textos escritos) o bien como una conversación, una conferencia, un discurso (de tratarse de textos orales). De esta manera, Rojo enfatiza en que se trata de textos construidos con intención comunicativa real y no concebidos para ilustrar un determinado fenómeno lingüístico.

El otro aspecto que Rojo (2021) analiza es el de la representatividad. Según él, los textos o fragmentos de textos contenidos en el corpus deben ser «conjuntamente representativos de una lengua o una variedad lingüística en un momento determinado de su historia o bien a lo largo de un cierto período» (p. 1). Pero, consciente de la complejidad que encierra la idea de representatividad, explica que como mínimo «el conjunto de textos integrados en un corpus debe dar una visión adecuada de aquello que pretende representar» (p. 1).

También analiza el carácter o naturaleza computacional del corpus. Rojo (2021) plantea que debido al tamaño de los corpus, a fin de poder recuperar la información necesaria para su estudio es ineludible pasar los textos a formato electrónico. Y afirma: «aunque conceptualmente pueda pensarse que el formato electrónico no es un rasgo constitutivo de la definición de los corpus, la realidad es que solo pueden ser manejados si poseen este carácter» (p. 2).

Por otro lado, es importante la codificación. Al respecto, Rojo (2021) señala que «los textos que forman un corpus deben estar codificados de modo que sea posible lograr la recuperación selectiva de la información» (p. 2). No se puede perder de vista que la selectividad es uno de los procedimientos metodológicos de la LC. Por eso, los textos se codifican: hay que añadir metadatos, es decir, una información paralingüística asociada al texto. La codificación, es decir, la inclusión de los metadatos en cada uno de los textos que componen un corpus, «hace posible su estudio científico y, más concretamente, la recuperación selectiva de la información que contiene» (p. 2) en la medida en que se puede efectuar esta extracción selectiva de datos valiéndose de una aplicación de consulta diseñada para tales efectos.

También la codificación implica la anotación lingüística. Rojo (2021) afirma que los textos que conforman un corpus

también se suelen someter a procesos de anotación lingüística, es decir, se le suelen añadir «una serie de informaciones referidas a sus características léxicas y gramaticales» (p. 2), porque «el análisis científico de un corpus incrementa considerablemente sus posibilidades si los textos que lo integran han sido sometidos también a un proceso de anotación lingüística» (p. 2).

Tras esta sistematización teórica de las definiciones de corpus, existe una última consideración que vale la pena añadir. Un corpus, como afirma Rojo «está formado por textos, pero es mucho más que un simple agregado de textos». En los mismos términos de este autor, «la palabra clave es diseño» (2021, p. 3). Por consiguiente, «la construcción de un corpus supone la integración sistemática de textos de acuerdo con un diseño determinado» (Rojo, 2021, pp. 23-24), lo que significa que «cada corpus tiene la configuración general que corresponde a los objetivos con los que se construye» (Rojo, 2021, p. 3). Esta es una de las razones por las que existen diferentes tipos de corpus y cada uno de ellos adecuado a su finalidad concreta.

Clasificación de los corpus lingüísticos

Como se concluye en el subepígrafe anterior, la construcción de un corpus responde a un diseño determinado de acuerdo con la finalidad con la que se construye. De ahí que exista una variedad de corpus que pudieran clasificarse atendiendo a diversas perspectivas. Entre las propuestas de tipología de clasificación de los corpus más destacadas en la bibliografía especializada sobresalen las de Torruella y Llisterri (1999), Procházková (2006), Villayandre (2008), Cruz (2017), Hincapié y Bernal (2018), Hincapié y Rubio (2018) y Rojo (2021).

Tras reflexionar sobre la tipologización de estos autores, sus criterios de clasificación y los tipos de corpus que incluyen dentro de esas categorías, se observa que algunos criterios de clasificación son muy generales y, por consiguiente, incluyen una gran variedad de tipos de corpus que no son equiparables ni opuestos; además, algunos de los corpus mencionados constituyen un subtipo o variante de otros enumerados en la misma lista.

Por otro lado, otros criterios clasificatorios resultan imprecisos porque, por ejemplo, parecen indicar cuestiones relativas al tamaño, pero en realidad presenta tipos de corpus que difieren más bien por el tamaño de la muestra seleccionada y por su finalidad y, es esta última la que condiciona la extensión y no de manera inversa. Hay que añadir, además, que algunas de las etiquetas clasificatorias podrían ser equívocas y por sí solas no pueden presentar la realidad que intentan describir.

Partiendo de lo analizado, se considera pertinente clasificar los corpus de acuerdo a los diferentes objetivos por los que se conciben y a su diseño, y en función de ello, teniendo en cuenta otra serie de propiedades como los tipos de textos que incluyen y sus características, el número de lenguas, el grado de información añadida, etc. Precisamente Rojo (2021) propone una tipología de corpus teniendo en cuenta estos presupuestos. Según este autor, los corpus pueden ser:

- *Corpus de referencia o especializados.* Estas etiquetas clasificatorias aluden a la finalidad con la que ha sido creado el corpus, criterio principal de clasificación. Los corpus de referencia, también denominados corpus generales, son los diseñados con la intención de ofrecer un recurso donde puedan ser analizados fenómenos y elementos lingüísticos que se producen en una determinada lengua, por tanto, intentan ser abarcadores de todo un dominio lingüístico y pueden estar compuestos por diferentes subcorpus. Por su parte, los corpus especializados son aquellos compuestos por la selección de textos con características comunes determinadas y pertenecientes a ámbitos muy variados. También son especializados los corpus de

aprendientes, constituidos por muestras orales o escritas de estudiantes con diferentes grados de dominio de una lengua extranjera.

- *Corpus orales y/o escritos.* Los corpus pueden contener textos orales y/o escritos. Estas etiquetas resultan evidentes: los corpus escritos son aquellos que contienen producciones escritas, mientras que los orales están conformados por intervenciones orales (y sus transcripciones). También existen los denominados corpus mixtos que contienen textos de ambas clases y más recientemente se habla de los multimodales.
- *Corpus totales o muestrales.* Un corpus total, por ejemplo, es aquel que contiene toda la obra de un autor o corriente literaria, todos los discursos pronunciados por un presidente, etc. Pero, como afirma Rojo (2021), estos corpus son de propósitos restringidos, razón por la cual son más utilizados dentro de la LC los corpus muestrales, los cuales se conciben como una muestra supuestamente representativa de una cierta lengua, variedad, etc. Por ejemplo, un corpus del español contemporáneo, de una variedad del español, del español de la prensa de un país y en un periodo determinado, etc.
- *Corpus cerrados o abiertos.* Un corpus cerrado es aquel que se planifica con un determinado tamaño, y que cuando se ha alcanzado el tamaño prefijado, se considera que está terminado y ya no se altera en su composición, a menos que se quiera añadir o modificar algún tipo de anotación lingüística. En cambio, un corpus abierto no parte con un tamaño ya establecido, sino que se concibe para que vaya creciendo a medida que lo hace posible la disponibilidad de textos y lo permiten las aplicaciones de explotación.
- *Corpus monolingüe o multilingüe.* Esta clasificación responde al criterio de selección de las lenguas de los textos que componen un corpus. Los corpus monolingües contienen textos que pertenecen solo a una lengua, en cambio, los multilingües se conforman con textos de más de una. Cabe destacar que estos últimos pueden presentarse de dos formas: corpus multilingües paralelos y corpus multilingües comparables. Los primeros contienen "el mismo texto" en dos o más lenguas, es decir, están formados por traducciones alineadas. Por el contrario, los corpus comparables están conformados por textos también en dos o más lenguas, pero sin ser traducciones unos de los otros.
- *Corpus sincrónicos o diacrónicos.* Si un corpus se centra en las características generales de una lengua en un momento determinado es de orientación sincrónica; en cambio, si trata de poner de manifiesto la variación existente en alguno de los ejes es de orientación diacrónica.
- *Corpus codificados o no codificados.* Atendiendo a la información añadida a los textos que permiten la recuperación de datos de manera selectiva, los corpus pueden estar codificados o no. Los codificados, los corpus más utilizados en la LC, actual añaden información extratextual (lugar, fecha, tipo de texto, etc.) congruente con la organización y diseño del corpus. Es decir, a los textos de estos corpus se les añade una cabecera con los llamados metadatos. Por el contrario, los no codificados no los añaden.
- *Corpus anotados y no anotados.* También según la información añadida a los textos que favorece la recuperación de datos de manera selectiva, los corpus pueden clasificarse como corpus anotados lingüísticamente y no anotados. Los primeros, los corpus más utilizados en la LC actual, son analizados en diferentes grados y niveles: fonológico, morfológico, lexical,

sintáctico, semántico, pragmático, etc. Esto es, los textos se someten a un proceso de anotación lingüística que favorece considerablemente el análisis científico.

Características de los corpus

En este subepígrafe se plantea identificar las características más importantes de un corpus lingüístico. Sin embargo, hay que enfatizar en que cualquier caracterización que se establezca podría resultar insuficiente ante la enorme cantidad de finalidades con que se puede concebir, diseñar y construir un corpus. En ese sentido, también hay que resaltar que la lista de características no debe poseer un matiz jerárquico, ni presentarse como una enumeración cerrada.

Se puede inventariar una gran lista de características importantes de los corpus lingüísticos, de hecho, existen algunas muy extensas, a saber, las ofrecidas por Torruella y Llisterri (1999), Bowker y Pearson (2002), Villayandre (2008), Parodi (2008) y Rojo (2021). Pero, como plantea Parodi (2008) «la descripción de un corpus radica de modo importante en la búsqueda de una especificación de sus características prototípicas» (p. 108), las cuales se consideran contenidas en la definición de corpus lingüísticos ofrecida por Rojo (2021) que ya se citó y comentó en el cuerpo de este texto y de la que se pueden desglosar rasgos prototípicos como:

- ser un conjunto de textos o fragmentos de textos;
- ser producidos en condiciones naturales;
- ser representativos de una lengua o una variedad lingüística, en su totalidad o en alguno(s) de sus componentes;
- ser almacenados en formato electrónico y
- ser codificados para que puedan ser analizados científicamente.

Así pues, aunque cada corpus lingüístico deberá cumplir aquellas características específicas congruentes con su objetivo concreto y diseño preestablecido, todos deberán respetar como mínimo estos rasgos prototípicos que resultan definitorios de lo que se entiende hoy en día como corpus dentro del campo de la LC.

Niveles de los corpus

Según su diseño, en un corpus lingüístico podemos hallar diferentes niveles. Una revisión de algunos con acceso abierto permite identificar como mínimo tres: corpus, subcorpus y componentes.

Precisamente, Torruella y Llisterri (1999) presentan esta clasificación de niveles jerarquizados. En un primer nivel sitúan al **corpus**, como conjunto de muestras de lengua de cualquier tipo que se toman como modelo de un estado o nivel de lengua predeterminado, que una vez analizado debe permitir mejorar el conocimiento del sistema lingüístico de la lengua que representa. En un segundo nivel ubican a los **subcorpus**, entendidos como una selección estática de textos, derivada de un corpus normalmente más general y complejo, el cual está dividido en grupos de muestras textuales más específicas; pero también puede ser una selección dinámica de textos de un corpus en crecimiento: un número determinado de textos destinados a aumentar algún apartado de un corpus general. Por último, sitúan a los **componentes**, comprendidos como una colección de muestras de un corpus o de un subcorpus, que responden a un criterio lingüístico específico muy concreto y, por consiguiente, reflejan un tipo determinado de lengua. Frente a los corpus y subcorpus que son heterogéneos, los componentes son muy homogéneos.

Corpus de aprendientes: hacia una definición y caracterización

Ya consolidada LC, en los años 90, surgen los corpus de aprendientes (Granger, 2002), un tipo de corpus especializado según la clasificación ofrecida por Rojo (2021). Una revisión bibliográfica permite constatar que no existe una estandarización a la hora de definir en

realidad en qué consisten estos corpus, y las definiciones que se encuentran (Sinclair, 1991; Granger, 2002; Hincapié, 2018; Calero, Serrano y Gómez, 2020; Lozano, 2022) suelen omitir rasgos que se consideran prototípicos de los corpus en general e incluso otros que caracterizan a los de aprendientes en particular.

Rojo (2021) explicita la finalidad de los corpus de aprendientes dentro de su tipología de corpus. Este autor plantea que estos corpus están constituidos por textos producidos por estudiantes de una determinada lengua segunda (L2) o extranjera (LE, con diferentes lenguas de origen y distintos grados de dominio de la L2/LE. Son corpus que se construyen precisamente para estudiar las características de la interlengua correspondiente y que, lógicamente, no pueden ser considerados como representativos de la L1 en cuestión. (p. 25)

En otras de sus obras de reciente publicación, Guillermo Rojo junto a Ignacio Palacios dedica un capítulo a los corpus de aprendientes de español como L2. En esta referencia Rojo y Palacios (2022) sostienen que un corpus de aprendientes consiste en la reunión de textos formulados por aprendientes de una (o varias) L2 y procedentes de una (o varias) L1, con la posibilidad de indicar también, entre los rasgos incorporados en la codificación, el nivel de conocimientos de la L2, el tipo de formación general, los años dedicados al estudio de la L2, el país de origen, etc. (p.75)

Seguidamente, exponen estos autores que, como todos los corpus, los de aprendientes «incorporan los textos tal como han sido producidos originalmente, con el mínimo de codificación necesario para la recuperación adecuada de datos» (Rojo y Palacios, 2022, p.75). Finalmente, agregan que a partir de ese mínimo de codificación, se pueden añadir rasgos adicionales de anotación morfosintáctica, concordancias, alineación de transcripción y sonido si se trata de muestras orales, codificación de errores, etcétera.

Sin embargo, si bien estas definiciones (Rojo, 2021; Rojo y Palacios, 2022) ofrecen luces sobre lo que es distintivo de los corpus de aprendientes, no presentan todos rasgos prototípicos que debe manifestar un corpus lingüístico según lo expuesto en el epígrafe anterior. Por eso, a partir de la definición general de corpus lingüístico formulada por Rojo (2021) comentada en este mismo texto y, teniendo en cuenta los rasgos distintivos de los corpus de aprendices en los que han enfatizado autores como Granger (2002), Hincapié (2018), Calero, Serrano y Gómez (2020), Rojo (2021) y Rojo y Palacios (2022), se propone la siguiente definición:

> Un corpus de aprendientes es un corpus especializado, constituido por un conjunto de textos, orales o escritos, que son reales, lo que implica que proceden de actos lingüísticos efectivamente realizados y producidos en condiciones naturales por estudiantes de una determinada lengua segunda (L2) o extranjera (LE), con una o diferentes lenguas maternas (L1) y con distintos grados de dominio de la L2/LE, que se almacena en formato electrónico y se codifica con metadatos y anotaciones lingüísticas con el fin de poder estudiar científicamente y con una orientación empírica las características de la interlengua correspondiente que se manifiestan a lo largo del proceso de aprendizaje.

Corpus de Aprendientes Chinos de Español: diseño, recolección y construcción

Objetivo: Construir un corpus informatizado, representativo de las variantes oral y escrita del español como lengua extranjera y segunda de los aprendientes chinos, para contar con una base de datos que permita llevar a cabo investigaciones rigurosas sobre su

interlengua e identificar las principales dificultades en su aprendizaje, a fin de diseñar estrategias lingüodidácticas adecuadas para su solución

Etapas en la construcción de CACHE

Para la construcción de CACHE se han establecido las tres etapas, resumidas en el Esquema 1 (Figura 1), y se han previsto acciones puntuales para cada una de ellas. A continuación, se describen.

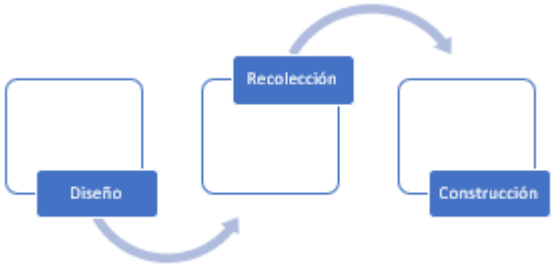

Fig. 1- Etapas planificadas para la construcción de CACHE

Etapa de diseño: Esta primera etapa se centra en el diseño y, por tanto, se corresponde con el periodo previo a la recogida de los datos. En ella, se han establecido las siguientes acciones:

1) Revisar de forma crítica la bibliografía especializada con vistas a sistematizar los referentes teóricos y metodológicos que fundamentan el diseño de un corpus.

2) Construir un diseño macro del corpus de aprendientes teniendo en cuenta la teoría sistematizada.

3) Asumir los principios teóricos y metodológicos en los que se cimenta el diseño del corpus que se construirá.

4) Determinar metadatos lingüísticos y extralingüísticos que se emplearán en la codificación y anotación lingüística de las muestras de lenguas que favorezcan la construcción de un corpus representativo de la interlengua española de los aprendientes chinos.

5) Crear y validar las tareas de expresión e interacción orales y escritas que se utilizarán como estímulos para la recogida de datos, teniendo en cuenta el objetivo general del corpus y los descriptores establecidos para estas destrezas en el Marco Común Europeo de Referencia y el Plan Curricular del Instituto Cervantes.

6) Diseñar y validar un instrumento informático adecuado para la recolección de los metadatos lingüísticos y extralingüísticos y la obtención de muestras de lengua, orales y escritas, que se incluirán en el corpus.

7) Determinar la herramienta lingüística informática que se empleará para la codificación lingüística y extralingüística y la anotación lingüística.

8) Determinar y diseñar el sistema informático con el que se gestionarán los datos y que constituirá la vía mediante la que se podrá acceder libremente al corpus por internet.

Etapa de recolección de datos: Esta etapa se corresponde con el período en que se recogen los datos. En ella, se ha contemplado básicamente la siguiente acción:

1) Recoger, a partir de las tareas e instrumentos creados y validados en la etapa de diseño, las muestras lingüísticas del español producidas por aprendientes chinos, teniendo en cuenta los objetivos, criterios de selección y los metadatos lingüísticos y extralingüísticos convenidos.

Etapa de construcción: Esta etapa se corresponde con el periodo en que se sistematizan todas las muestras de lengua obtenidas y se finaliza la construcción del corpus. En ella se contemplan las siguientes acciones:

1) Seleccionar para su inclusión en el corpus aquellas muestras de lenguas que cumplan los criterios de selección convenidos y desechar las restantes.

2) Efectuar, automática y manualmente, el proceso de codificación y anotación de las muestras aprobadas para su inclusión en el corpus, a partir de los metadatos y etiquetas convenidas.

3) Construir la herramienta informática de búsqueda del corpus, una Web, que permita recuperar información estadística y lingüística; y publicarla con acceso abierto para uso de la comunidad científica.

4) Publicar el corpus, según la vía determinada, con acceso abierto para uso de la comunidad académica y científica.

Ejecución de la etapa de diseño

Tras cumplir con la primera tarea de la etapa de diseño (la revisión crítica de la teoría y metodología referidas al diseño de los corpus) se prosiguió a la ejecución de las restantes acciones previstas para este periodo. Seguidamente, se exponen los aspectos más significativos del diseño de CACHE, resultado del cumplimiento de dichas tareas.

Generalidades del diseño macro

El corpus de aprendientes que se pretende construir se denominará Corpus de Aprendientes Chinos de Español (CACHE). Una vez construido, será un corpus especializado, muestral y representativo, multilingüe -contemplará un corpus nativo de control en español y chino mandarín-, mixto -incluirá muestras orales y escrita-, y abierto; además, será de orientación transversal.

Como se observa en el Esquema 2 (Figura 2), CACHE estará conformado por cuatro subcorpus: dos subcorpus de interlenguas de aprendientes de español y los dos subcorpus de control en español y chino mandarín.

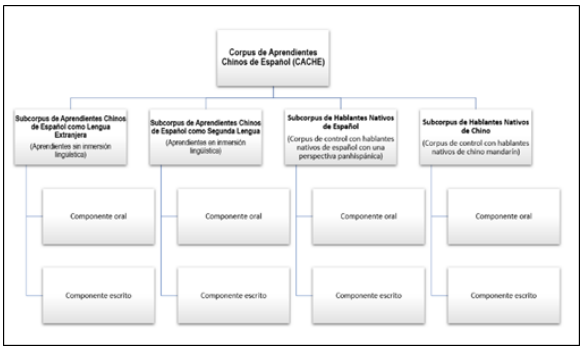

Fig. 2- Diseño macro de CACHE

El núcleo del corpus lo constituyen los dos subcorpus de aprendientes: el Subcorpus de Aprendientes Chinos de Español como Lengua Extranjera, en el que participarán aprendientes sin inmersión lingüística, y el Subcorpus de Aprendientes Chinos de Español como Segunda Lengua, en el que colaborarán aprendientes en contextos de inmersión. Ambos subcorpus incluirán dos niveles: un componente oral y otro escrito. También se compondrá de dos subcorpus de control, de la lengua hablada y escrita por los nativos de español y del chino mandarín. Se incluirá un Subcorpus de Hablantes Nativos de Español, en el que se recogerán muestras de lengua de hablantes nativos de español (de cualquiera de las comunidades lingüísticas del mundo hispano); y un Subcorpus de Hablantes Nativos de Chino, en el que se recopilarán muestras de lengua de hablantes de chino mandarín.

Principios teóricos y metodológicos para la fundamentación del diseño

El diseño de un corpus cualquiera debe fundamentarse en criterios teóricos y metodológicos específicos. Se considera que los principios de creación de un corpus formulado por Sinclair (2005) son válidos como fundamento principal del diseño de los corpus de aprendientes. La creación de CACHE se rige por los diez principios de creación de corpus por dicho autor que se describen a continuación:

Principio para la selección del contenido: «The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise» (Sinclair, 2005, p. 1). Los textos que conformarán CACHE serán producidos en condiciones naturales, por lo que se crearán según criterios externos, es decir, en dependencia de su función comunicativa. No se considerarán criterios internos referidos a la lengua del texto, es decir, con el objetivo de recoger una determinada estructura lingüística.

Principio de la representatividad: «Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen» (Sinclair, 2005, p. 2). CACHE no puede ser un corpus representativo del español, porque se trata de un corpus de aprendientes. Este corpus será representativo de la interlengua de los aprendientes chinos de español. Ahora bien, cuando se habla de representatividad aquí, no se pretende representar la "totalidad" de la lengua española producida por aprendientes chinos, pues un corpus que pretenda eso, tendría que recoger datos, por ejemplo, en todos los centros educativos donde se aprende español en China y donde los chinos aprenden español en el extranjero. Así pues, dicha representatividad se logra con la inclusión de muestras de interlengua en diferentes estadios de desarrollo toda vez que en su construcción participarán aprendientes de los niveles A1- C2. De hecho, CACHE constituirá un corpus a partir de un muestreo de orientación transversal que considerará los distintos niveles de competencia. Por otro lado, con el fin de obtener una variedad de estructuras lingüísticas, se ofrecerán diferentes tareas comunicativas a modo de estímulos para las producciones textuales y se tendrán en cuenta la diversidad de tipos de textos, otros aspectos que prevé la representatividad toda vez que la diversidad de temas y tipos de textos favorece la elicitación variadas formas y estructuras lingüísticas.

Principio del contraste: «Only those components of corpora which have been designed to be independently contrastive should be contrasted» (Sinclair, 2005, p. 3). Lo descrito en el segundo principio favorece el cumplimiento del tercero, porque la representatividad con la inclusión de muestras de interlengua en diferentes estadios de desarrollo con la participación de aprendientes de los niveles A1- C2, permite contrastar diferentes estadios de la interlengua. Pero, además, el diseño de CACHE comprende la conformación de dos subcorpus de interlengua: un Subcorpus de Aprendientes Chinos de Español como Lengua Extranjera donde participarán aprendientes en

contextos de no inmersión lingüística y un Subcorpus de Aprendientes Chinos de Español como Segunda Lengua donde participarán aprendientes en contextos de inmersión lingüística. Cada uno de estos subcorpus dispondrá de un componente oral y escrito. El diseño, la metodología, y las tareas previstas para la recogida de los datos serán iguales para ambos subcorpus, de modo que quede garantizada la posibilidad de realizar contrastes de interlengua entre aprendientes en contexto de no inmersión e inmersión lingüística. Además, de la misma manera se concibe la recogida de datos del Subcorpus de Hablantes Nativos de Español y del Subcorpus de Hablantes Nativos de Chino, lo que permite hacer contrastes fiables. Como afirma Cruz (2017), «el análisis empírico de un corpus de aprendices generalmente se realiza comparándolo con un "corpus de control" que está constituido por textos de características similares (tema, tipo de texto, etc.) producidos por nativos». (p. 138). El poder contrastar la interlengua de los estudiantes chinos con la lengua producida por nativos del español, posibilita el logro de una mayor objetividad en el análisis.

Principio para la determinación de los criterios estructurales: «Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination» (Sinclair, 2005, p. 5). Como CACHE intenta ser un corpus grande, este criterio resulta muy importante. Ahora bien, no se puede aplicar este criterio de la manera en que se aplicaría a los extensos corpus de referencia. CACHE es un corpus de aprendientes y su diseño estructural marco, descrito en el tercer principio, ya se rige por criterios exactos: según el contexto de estudio de los aprendientes (con sin inmersión lingüística o sin ella) se estructura en dos subcorpus que, a su vez, y teniendo en cuenta el medio de producción de los textos (orales o escritos) se dividen en dos componentes. Además, el diseño micro contemplará una división por los niveles de lenguas A1-C2.

Principio para el establecimiento del etiquetado: «Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications» (Sinclair, 2005, p. 5). Para cumplir con este principio, cuando se vaya a comenzar con el proceso de codificación de los textos de CACHE, la inclusión de metadatos y la anotación lingüística, se seleccionará una herramienta lingüística que tras el proceso de codificación produzca un nuevo fichero XML con el texto con metadatos y anotaciones y conserve el

fichero original con el texto limpio.

Principio para la determinación de la muestra: «Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events or should get as close to this target as possible. This means that samples will differ substantially in size» (Sinclair, 2005, p. 7). En CACHE solo se incluirán textos completos sin considerar su extensión en número de palabras, porque, como enfatiza Sinclair (2005), seleccionar muestras del mismo tamaño no posee ningún valor desde el punto de vista lingüístico, pero la integridad y la representatividad de los textos completos si adquiere alto valor.

Principio para la selección de la documentación: «The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken» (Sinclair, 2005, p. 8). Dado que CACHE es un corpus que pretende ser una fuente de datos de acceso libre con la que numerosos lingüistas y profesores de español podrán desarrollar variados estudios sobre la interlengua de los estudiantes chinos que estudian español, la Web que lo albergará contendrá todos los documentos que fundamentan teórica y metodológicamente el diseño y la recogida de datos, información relativa a los metadatos incluidos en el corpus y al proceso de anotación lingüística, etc.

Principio del equilibrio: «The corpus builder should retain, as target notions, representativeness, and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components» (Sinclair, 2005, p. 9). En este sentido, el interés por la representatividad y el equilibro está explícito en el propio diseño de CACHE que contiene un componente escrito y uno oral en cada uno de sus subcorpus. De esta forma, los resultados obtenidos a partir del uso de CACHE podrán ser extrapolados tanto a la interlengua escrita como oral.

Principio para la determinación del tema: «Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria» (Sinclair, 2005, p.10). En relación con este, el cumplimiento del primer principio referido al contenido del corpus y del segundo que exige la representatividad, obliga a regirse por el noveno. Los controles que se efectuarán estarán relacionados con criterios externos como el nivel de interlengua, pero nunca se aplicarán criterios que condicionen las muestras de lengua en busca de unas estructuras lingüísticas concretas.

Principio de la homogeneidad: «A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided» (Sinclair, 2005, p. 14). CACHE desde su diseño considera el respeto a la homogeneidad en sus componentes y plantea unos criterios claros para evitar la introducción de textos atípicos en el corpus. Por ejemplo, quedarán excluidos aquellos textos que no constituyan una unidad de comunicación completa y otros que presenten señales de empleo de traducción.

Metadatos lingüísticos y extralingüísticos y anotación

Con reiteración se ha hablado, en estas páginas, de la objetividad que se logra con la investigación lingüística basada en corpus; se ha declarado que la lingüística de corpus es, precisamente, una aproximación empírica. Siempre se ha considerado que esta objetividad se alcanza con la construcción de un corpus representativo de la realidad lingüística en cuestión; sin embargo, este grado de objetividad no responde solamente a la exhaustividad, también depende, y en gran medida, del empleo de metadatos lingüísticos y extralingüísticos adecuados, garantes también de la representatividad y de la posibilidad de contrastes. Así pues, CACHE contará con un conjunto de metadatos bien definidos que favorecerá la explotación del corpus, pues todos los empleados en la codificación y anotación permitirán filtrar información en las búsquedas automáticas.

Como se ha expresado con anterioridad, al solo tenerse en cuenta en esta investigación el componente escrito del Subcorpus de Aprendientes Chinos de Español como Lengua Extranjera, el del Hablantes Nativos de Español y el del Subcorpus de Hablantes Nativos de Español, únicamente se han determinado los metadatos lingüísticos (textuales) y extralingüísticos (sociolingüísticos) que se usarán en la codificación de los textos de estos subcorpus.

Metadatos lingüísticos

Como se aprecia en la tabla 1, los metadatos lingüísticos que se consideran para la codificación de los textos reflejan información importante relacionada con el género y tipo de texto, el tipo de tarea, la cantidad de palabras y párrafos y el uso de diccionarios, además, indican si el texto fue escrito en un contexto controlado (en el aula, bajo supervisión de profesores, etc.) o no y la identidad de los encargados de examinar los textos para determinar su inclusión o exclusión del corpus.

Tabla 1- Metadatos lingüísticos para la codificación

Metadatos	Descripción
Id de la muestra	SUBACHELE_CE_001 SUBHNE_CE_001 (Nombre del subcorpus + componente escrito + número de la muestra)
Género textual predominante	Expositivo / descriptivo / narrativo / argumentativo
Tipo de texto	Carta / correo / cuento / biografía / artículo de opinión / reseña crítica / mensaje (blog /foro), etc.
Tipo de tarea	Expresión escrita / expresión e interacción escrita
Número de palabras	-
Número de párrafos	-
Empleo de diccionario	Sí / No
Cantidad de búsquedas efectuadas	-
Controlado	Sí / No
Primer revisor	Nombre (siempre sería la autora)
Segundo revisor	Nombre (siempre sería un especialista nativo de español)

Metadatos extralingüísticos

Por su parte, en la tabla 2, se relacionan los metadatos extralingüísticos que se tienen en cuenta en la codificación de los textos. Entre ellos, destacan las variables sociales habituales en la sociolingüística (sexo, edad, nivel de instrucción, lengua materna), consideradas variables adscritas. Pero también, se incluyen otras variables denominadas adquiridas como las lenguas extranjeras, el nivel adquirido en esas lenguas extranjeras, las estancias en el extranjero, los estudios y centros de estudios, y la motivación ante el estudio del español.

Tabla 2- Metadatos extralingüísticos para la codificación según la información de los informantes

Metadatos	**Descripción**
Id del aprendiente	SUBACHELE_CE_001_H1A1 (Id de la muestra + sexo (H/M) + nivel de instrucción (1/2) + nivel en español (A1-C2)) SUBHNE_CE_001_M2 (Id de la muestra + sexo (H/M) + nivel de instrucción (1/2))
Sexo	Hombre (H) Mujer (M)
Edad	
Sobre el nivel de instrucción	
Estudios terminados	Preuniversitario (P) / Grado (G) / Máster (M) / Doctorado (D)
Estudios en curso	Grado (G) / Máster (M) / Doctorado (D)
Carrera / Posgrado	Titulación
Centro de estudio	Nombre
Año que cursa	
Sobre la lengua	
Lengua materna (L1)	Chino mandarín / Español /Otra ¿Cuál?
Segunda lengua materna	Otras lenguas chinas, geolectos o dialectos aprendidos en el seno de la familia
10 lengua extranjera/segunda lengua	Inglés / japonés / coreano /otras ¿Cuál?
Autovaloración del nivel estimado en esa lengua (Subjetivo)	-
Nivel en esa lengua (Objetivo-Certificaciones)	Certificaciones internacionales
20 lengua extranjera/segunda lengua	Inglés / japonés / coreano /otras ¿Cuál?
Sobre el estudio del español (solo para los informantes chinos)	
Autovaloración del nivel estimado en español (Subjetivo)	A1/ A2 / B1/ B2/ C1/ C2
Nivel de español (Objetivo-Certificaciones)	EEE4 /8; CATTI 3 / 2 / 1; SIELE A1/ A2 / B1/ B2/ C1; DELE A1/ A2 / B1/ B2/ C1/ C2
Nivel de español (Objetivo-Prueba de nivel)	Resultado de la prueba de nivel
Tiempo aprendiendo español	Meses o años
Estancia de estudio o intercambio en un país hispano	Sí / No
País hispano	Nombre del país
Tiempo de estancia en ese país	Meses o años
Estancia de trabajo en un país hispano	Sí / No
País hispano	Nombre del país
Tiempo de estancia en ese país	Meses o años
Motivación / Actitud ante el estudio del español (Autoevaluación)	1-10

Algunos de estos metadatos se emplearán para la codificación de los textos en la web del corpus y servirán para filtrar información en búsquedas automáticas, por ejemplo, la edad, el nivel de idiomas, etc. Sin embargo, otros serán de utilidad para desambiguar ante cualquier duda que aparezca cuando se examinen los datos. Así, por ejemplo, ante un texto de un informante que cursa el tercer año de la carrea en Filología Hispánica y lleva unos dos años estudiando español y no posea ninguna certificación que acredite su nivel de idioma, se contrastará el nivel de lengua con el que se ha autoevaluado con el nivel que señale la prueba de nivel, y se tendrá en cuenta el año de la carrera que cursa. De esta manera, se podrá ubicar mejor en el nivel de lengua que le corresponde. En suma, algunos de estos datos serán usados en el diseño del corpus, mientras que otros solo permitirán hacer una mejor evaluación de los textos recogidos.

CONCLUSIONES

La revisión crítica de la teoría disponible en la bibliografía especializada acerca de la lingüística del corpus, los corpus lingüísticos y, en particular, los corpus de aprendientes, permitieron sistematizar a lo largo de estas páginas las nociones teóricas y metodológicas fundamentales y proponer una definición precisa de corpus de aprendientes. De modo que, la primera contribución de esta investigación se sitúa en el orden teórico con la sistematización teórica puesta a disposición de futuros estudios en esta área.

Otra contribución teórica y también práctica se manifiesta en la fundamentación teórica y metodológica del diseño del Corpus de Aprendientes Chinos de Español y en la determinación de las etapas para su creación: diseño, recolección de datos y construcción. Asimismo, se exhiben los resultados alcanzados en la ejecución de la etapa de diseño: la concepción del corpus a partir de sus objetivos, sus generalidades, sus subcorpus y componentes, los principios teóricos y metodológicos en los que se fundamenta su creación y los metadatos lingüísticos y extralingüísticos para su codificación.

El Corpus de Aprendientes Chinos de Español (CACHE), una vez construido, será un corpus especializado, muestral y representativo, multilingüe, mixto, abierto y de orientación transversal. Estará accesible a la comunidad científica con el fin de fomentar el desarrollo de investigaciones que contribuyan al conocimiento de las características de la interlengua que producen los estudiantes chinos de español como lengua extranjera/segunda lengua. Se continúa trabajando en las etapas de recolección de datos y su proceso construcción.

REFERENCIAS BIBLIOGRÁFICAS

Aarts, J. (2002). Does Corpus Linguistics Exist? Some Old and New Issues. Egil, B. L. y Hasselgren, A. (eds). *From the COLT's Mouth ... And Others' Language Corpora Studies. In Honour of Anna-Brita Stenström - Language and Computers* (pp. 117). Editions Rodopi B.V. https://brill.com/display/title/30088

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, 8*(4), 243-257. https://doi.org/10.1093/llc/8.4.243

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511804489

Bolaños, S. (2015). La lingüística de corpus: perspectivas para la investigación lingüística contemporánea. *Forma y Función*, 28, 1, 31-54. https://doi.org/10.15446/fyf.v28n1.51970

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language A Practical Guide to Using Corpora* (1era ed.). Routledge. https://www.routledge.com/Working-with-Specialized-Language-A-Practical-Guide-to-Using-Corpora/Bowker-Pearson/p/book/9780415236997

Calero F., Mª A.; Serrano Z., Mª I.; Gómez-Devís, Mª 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. https://cvc.cervantes.es/lengua/eaesla/eaesla_06.htm

Cruz, M. (2012). *Lingüística de corpus y enseñanza del español como 2/L.* Arco Libros. https://doi.org/10.5565/rev/doble.35

Cruz Piñol, M. (2017). *Lingüística de corpus y enseñanza del español como 2/L* (2da ed.). Editorial La Muralla. https://www.arcomuralla.com/detalle_libro.php?id=872&ideditorial_get=1

Francis, W. N., Kucera, H., & Mackie, A. W. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Houghton Mifflin. https://lib.ugent.be/en/catalog/rug01:000049253

Granger, S. (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. En Lllt.6. John Benjamins Publishing Company. https://benjamins.com/catalog/lllt.6

Gries, S. T., & Stefanowitsch, A. (2006). Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis. Mouton de Gruyter. https://books.google.com.pe/books?id=D_nVI5mFtkAC&lr=&num=20

Gries, S. (2009). What Is Corpus Linguistics. *Language and Linguistic Compass* 3, 12251241. https://doi.org/10.1111/j.1749-818X.2009.00149.x

Hincapié, D. (2018). Corpus de aprendientes de español como lengua extranjera y segunda lengua (CAELE/2): el componente escrito. *Forma y Función*, 31(2), 129-144. https://doi.org/10.15446/fyf.v31n2.74659

Hincapié M., D. y Bernal C., J. (2018). *Lingüística de corpus.* Instituto Caro y Cuervo. https://bibliotecadigital.caroycuervo.gov.co/1703/1/Linguistica-de-corpus-2018.pdf

Hincapié, D. y Rubio, R. (2018). Diseño y construcción del CAELE2: Base para una planificación curricular. *Hechos y Proyecciones del Lenguaje*, 23 (1), 42 52. https://revistas.udenar.edu.co/index.php/rheprol/article/view/3842

Krug, M., Schlüter, J., & Rosenbach, A. (2013). Introduction: Investigating language variation and change. In M. Krug & J. Schlüter (Eds.), *Research Methods in Language Variation and Change* (pp. 1-14). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511792519

Leech, G. (1992). Corpora theories of linguistic performance. En Svartvik, J. (Ed.). *Directions in Corpus Linguistics*. (pp. 105-122). Mouton de Gruyter. https://doi.org/10.1515/9783110867275

Leech, G (2011). Principles and Applications of Corpus Linguistics. En Viana, V., Zyngier, S. y Barnbrook, G. *Perspectives on Corpus Linguistics (Studies in Corpus Linguistics, 48)* (pp. 155170). John Benjamins Publishing Company. https://doi.org/10.1075/scl.48

Lemnitzer, L., & Zinsmeister, H. (2008). Korpuslinguistik. Eine Einführung. 2. Auflage. Tübingen: NarrFrancke Attempto Verlag. https://doi.org/10.1515/infodaf-2008-2-362

Lozano, C. (2022). CEDEL2: Diseño, compilación e interfaz web de un corpus online para la investigación de adquisiciones de L2 en España. *Investigación en un segundo idioma*, *38*(4), 965-983. https://doi.org/10.1177/02676583211050522

McEnery, A. M., Xiao, R. Z. and Tono, Y. (2006). *Corpus-based language studies: an advanced resource book.* Routledge Applied Linguistics Series. Routledge. https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm

Parodi, G. (2008). Lingüística de corpus: Una introducción al ámbito. *RLA. Revista de lingüística teórica y aplicada*, 46(1), 93-119. https://doi.org/10.4067/S0718-48832008000100006

Procházková, P. (2006). Fundamentos de la lingüística de corpus. Concepción de los corpus y métodos de investigación con corpus. https://docplayer.es/78249763-Fundamentos-de-la-linguistica-de-corpus.html

Rojo, G. (2022). *Introducción a la lingüística de corpus en español. LinRed*, (XIX). https://erevistas.publicaciones.uah.es/ojs/index.php/linred/article/view/1974

Rojo, G. y Palacios, I. (2022). Los corpus de aprendientes de español como L2. En: Parodi, G., Cantos-Gómez, p., Howe, C. (Eds). *Lingüística de corpus en español, (pp. 73-88). The Routledge Handbook of Spanish Corpus Linguistics.* https://www.routledge.com/Linguistica-de-corpus-en-espanol—The-Routledge-Handbook-of-Spanish/Parodi-Cantos-Gomez-Howe/p/book/9780367350123

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press. https://repositorio.uchile.cl/bitstream/handle/2250/139995/Corpus-concordance-collocation.pdf?sequence=4

Sinclair, J., & Wynne, M. (2005). How to build a corpus. En *Developing Linguistic Corpora: A Guide to Good Practice* (p. 108). AHDS. http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf

Torruella, J. y Llisterri, J. (1999) Diseño de corpus textuales y orales. En Blecua, J.M., Clavería, G., Sánchez, C., Torruella, J. (Eds.) *Filología e informática. Nuevas tecnologías en los estudios filológicos.* Universidad Autónoma de Barcelona. (pp. 45-77). Milenio. https://dialnet.unirioja.es/servlet/articulo?codigo=595883

Tognini-Bonelli (2001). *Corpus Linguistics at WorK.* John Benjamins Publishing Company. https://doi.org/10.1075/scl.6

Villayandre Llamazares, M. (2008). Lingüística con corpus (I). *Estudios humanísticos. Filología*, *30*, 329-349. https://doi.org/10.18002/ehf.v0i30.2847

Conflicto de intereses:

Los autores declaran no tener conflictos de intereses.

Contribución de los autores:

Los autores participaron en el diseño, análisis de los documentos y redacción del trabajo.

Citar como
Gou, J., Rodríguez Roque, D., Cuba Vega, L.E. (2023). Diseño teórico y metodológico del Corpus de Aprendientes Chinos de Español. *Mendive. Revista de Educación, 21*(4), e3347. https://mendive.upr.edu.cu/index.php/MendiveUPR/article/view/3347