

## CORPUS SUBCACHELE ESCRITO: ESTUDIO DE LA INTERLENGUA DE LOS ESTUDIANTES CHINOS Y SUS APLICACIONES DIDÁCTICAS

*SUBCACHELE Written Corpus: A Study of the Interlanguage of Chinese Students and Its Didactic Applications*

*Corpus SUBCACHELE Escrito: um estudo da interlíngua de estudantes chineses e suas aplicações didáticas*

**Jialing Gou**

wawaana0601@gmail.com

<https://orcid.org/0000-0002-0303-974X>

重庆外语外事学院 [Chongqing Institute of Foreign Studies] (China)

Recibido: 11/07/2024

Revisado: 10/10/2024

Evaluado: 17/05/2025

Aceptado: 19/05/2025

268

### Resumen

En la última década, ha aumentado el interés por los corpus de aprendientes de español como lengua extranjera (ELE). Sin embargo, al revisar los corpus accesibles, constatamos que no se han recopilado muestras exhaustivas de interlengua española de los estudiantes de Filología Hispánica en China. Dentro de la lingüística de corpus aplicada a la enseñanza de ELE, describimos el componente escrito del Subcorpus de Aprendientes Chinos de Español como Lengua Extranjera (SubCACHELE Escrito), que forma parte del corpus CACHE, y reflexionamos sobre sus aplicaciones. Para su construcción, entre 2022 y 2024 recopilamos textos escritos por estudiantes chinos de los cuatro años de Filología Hispánica de varias universidades chinas. Participaron 567

informantes, de ambos sexos y con una media de edad de 21 años. En este trabajo, describimos tecnológica y estadísticamente el SubCACHELE Escrito y sus potencialidades para la investigación de la interlengua de estos estudiantes y como recurso didáctico para la enseñanza universitaria del español en China.

## Abstract

In the last decade, there has been increased interest in learner corpora of Spanish as a Foreign Language (SFL). However, upon reviewing the accessible corpora, we found that comprehensive samples of the Spanish interlanguage among Chinese students majoring in Hispanic Philology had not been collected. Within corpus linguistics applied to SFL teaching, we describe the written component of the Subcorpus of Chinese Learners of Spanish as a Foreign Language (SubCACHELE Written), which is part of the CACHE corpus, and reflect on its applications. For its construction, between 2022 and 2024, we collected written texts from Chinese students in all four years of Hispanic Philology from various Chinese universities. A total of 567 informants participated, of both genders, with an average age of 21 years. In this work, we describe the SubCACHELE Written technologically and statistically, as well as its potential for researching these students' interlanguage and its use as an educational resource for university-level Spanish teaching in China.

269

**Palabras Clave:** corpus de aprendientes, español como lengua extranjera, subcorpus de escritura, Filología Hispánica, estudiantes chinos.

**Keywords:** Learner Corpus, Spanish as a Foreign Language, Writing Subcorpus, Hispanic Philology, Chinese Students.

## Introducción

En este trabajo, además de difundir una parte de los resultados obtenidos durante la investigación doctoral titulada *Diseño de un corpus de aprendientes chinos de español: potencialidades para el estudio y el tratamiento de la*

*atenuación en español*, se describe la labor que hemos realizado en los últimos años para construir el Corpus de Aprendientes Chinos de Español (CACHE), y en particular, el componente escrito del Subcorpus de Aprendientes Chinos de Español como Lengua Extranjera (SubCACHELE Escrito).

En Gou Jialing, Rodríguez Roque y Cuba (2023) expusimos cómo surgió el interés por la creación de este subcorpus tras constatar que los corpus de aprendientes de ELE existentes y accesibles en línea no incluían muestras representativas de la interlengua española de los estudiantes de Filología Hispánica en China. En ese trabajo precedente, nos referimos al estado de la cuestión relativo a los corpus de aprendientes de español, con una descripción detallada de las muestras de interlengua de los estudiantes chinos de español presentes en los corpus consultados, lo que nos permitió confirmar la necesidad del presente estudio.

## Objetivo

Dentro de la lingüística de corpus aplicada a la enseñanza de ELE en China, en este trabajo perseguimos como objetivo general describir el componente escrito del Subcorpus de Aprendientes Chinos de Español como Lengua Extranjera (SubCACHELE Escrito), que forma parte del Corpus de Aprendientes Chinos de Español (CACHE), y reflexionar sobre sus aplicaciones en la investigación científica y en la enseñanza dentro del contexto universitario chino.

### *Tareas de investigación:*

- Revisar la teoría que sustenta la creación de un corpus de aprendientes, lo que implica identificar o desarrollar una definición que reúna los rasgos prototípicos de este tipo de corpus.
- Ratificar los principios (asumidos en Gou Jialing, Rodríguez Roque y Cuba, 2023) que deben guiar la creación de un corpus de aprendientes como el SubCACHELE Escrito, lo que conlleva la puesta en práctica de los procedimientos metodológicos sobre los que se sustenta la creación de este corpus.

- Desarrollar las etapas de diseño, recopilación y creación del SubCACHELE Escrito, previstas desde la concepción de su diseño (Gou Jialing, Rodríguez Roque y Cuba, 2023)
- Recopilar la estadística del SubCACHELE Escrito y presentarla a modo de resultados y como referencia para la toma de decisiones de los investigadores interesados en esta fuente de datos empíricos.
- Valorar posibles aplicaciones del subcorpus creado, lo que incluye identificar áreas de aplicación en la investigación científica y la enseñanza de ELE.

## Método

### *Etapas de diseño, recopilación y construcción*

Una definición no parece suficiente para comprender en qué consiste un corpus de aprendientes. Para entenderlo, además de definirlo, hay que aceptar que «siempre responden a un diseño previo» (Rojo, 2021, p. 81). Por ello, primeramente, concebimos el diseño teórico-metodológico del Corpus de Aprendientes Chinos de Español (CACHE). En Gou Jialing, Rodríguez Roque y Cuba (2023) describimos las bases teóricas y metodológicas del diseño del corpus y recogimos las tres etapas (diseño, recopilación de datos y construcción) establecidas para su creación y las acciones puntuales previstas para cada una de ellas. Seguidamente, retomamos las ideas más relevantes.

La etapa de diseño, que se corresponde con el periodo previo a la recogida de los datos, incluyó la sistematización de los referentes teóricos y metodológicos para el diseño de un corpus de aprendientes; la elaboración de un diseño macro; la elección de los principios teóricos y metodológicos, cimientos de la construcción del corpus; la determinación de metadatos (extra)lingüísticos para la codificación y anotación lingüística de los datos; la creación y posterior validación de las tareas para la recogida de datos; la validación del instrumento para la recolección de datos a través de una prueba piloto; la determinación de la herramienta lingüística informática para la codificación y la anotación y el diseño del sistema informático para gestionar los datos y ponerlos accesibles (Gou Jialing, Rodríguez Roque y Cuba, 2023).

Tras esta etapa de diseño, cuyas acciones fueron fundamentales, se pasó a la etapa de recolección de datos, que contemplamos básicamente para la recogida de las muestras lingüísticas producidas en idioma español por los aprendientes chinos. Por su parte, en la etapa de construcción, se incluyó la selección de las muestras de lenguas que cumplieran los criterios establecidos para su inclusión en el corpus; la codificación automática y manual, a partir de los metadatos y etiquetas establecidas, de las muestras incluidas en el corpus; la construcción de un sitio web para la recuperación de información estadística y lingüística y, por último, su publicación para su uso público (Gou Jialing, Rodríguez Roque y Cuba, 2023).

#### *Principios teóricos y metodológicos para la fundamentación del diseño*

En Gou Jialing, Rodríguez Roque y Cuba (2023) fundamentamos teórica y metodológicamente la creación del corpus CACHE en los principios formulados por Sinclair (2005). La construcción del SubCACHELE Escrito también se rige por estos diez principios: el principio de la selección del contenido, el de la representatividad, el de contraste, el de la determinación de los criterios estructurales, el del establecimiento del etiquetado, el de la determinación de la muestra, el de la selección de la documentación, el de equilibrio, el de la determinación del tema y el principio de la homogeneidad (Sinclair, 2005).

272

#### *Instrumento, recogida de datos e informantes*

Entre 2022 y 2024, para la construcción del SubCACHELE Escrito, apoyados en la lingüística de corpus y en la fundamentación teórica y metodológica del diseño previo (Gou Jialing, Rodríguez Roque y Cuba, 2023), recopilamos, mediante un cuestionario en línea, textos escritos por 567 estudiantes chinos, de ambos sexos, con una media de edad de 21 años y de los cuatro años de Filología Hispánica de universidades chinas como Dalian Foreign Languages University (Dalian), Chengdu International Studies University (Chengdu), Chongqing Institute of Foreign Studies (Chongqing), Shanghai International Studies University (Shanghai) Sichuan International Studies University (Chongqing) y Xi'an International Studies University (Xi'an).

Para la recolección de datos, elaboramos un cuestionario en Microsoft Forms, que permitió recopilar información y exportarla a Microsoft Excel. El cuestionario está formado por tres secciones: (1) información general de la investigación y consentimiento informado, (2) metadatos extralingüísticos determinados para la codificación de los textos y (3) tareas de escritura para la redacción de los textos del SUBCACHELE Escrito.

En cuanto a las tareas de escritura, creamos un conjunto de diez para cada grupo de niveles A (A1 y A2), B (B1 y B2) y C (C1 y C2), en cada caso se presentaron cinco tareas de expresión escrita y cinco de interacción escritas, teniendo en cuenta los géneros textuales y tipos de textos fijados en el Plan Curricular del Instituto Cervantes para los niveles de referencia del español, que a su vez sigue las recomendaciones del MCER. De este modo, garantizamos la variedad de géneros (exposición, descripción, narración, argumentación) y tipos de texto (mensajes, correos, cartas personales y profesionales, artículos de opinión, informes, reseñas críticas, etc.), al igual que la variedad de ámbitos temáticos (personal, público, profesional, académico), lo que nos permitió obtener muestras variadas de la interlengua de los informantes.

También fue necesaria la validación del cuestionario. Para ello, llevamos a cabo una prueba piloto en el segundo semestre del curso 2021-2022, en la que participaron estudiantes de la Chengdu International Studies University (Chengdu). Tras esta prueba, analizamos los resultados y recogimos opiniones sobre la comprensión del cuestionario. Como resultado, tomamos las siguientes acciones: añadimos aclaraciones en algunas preguntas y conformamos una versión bilingüe (español y chino) del cuestionario, a fin de garantizar la fiabilidad en la obtención de los metadatos extralingüísticos. Cabe señalar que la tarea de redacción se presentó exclusivamente en español.

## Marco teórico

### *Definición, caracterización y tipología de corpus lingüísticos*

En Gou Jialing, Rodríguez Roque y Cuba (2023) revisamos las definiciones de corpus lingüísticos y corpus de aprendientes más recurrentes en la bibliografía especializada. Para ese trabajo examinamos las propuestas de Francis, Kučera y Mackie (1982), Leech (1991, 1992), Sinclair (1991, 1996, 2005), Biber (1993), Biber *et al* (1998), Torruella y Llisterri (1999), McEnery (2001), Santalla del Río (2005), McEnery, *et al* (2006), Villayandre (2008), Mercado (2008), Parodi (2008), Lemnitzer y Zinsmeister (2008), Hincapié y Rubio (2018), Hincapié y Bernal (2018) y la de Rojo (2021). Esta revisión nos permitió analizar las tendencias históricas de las definiciones de corpus lingüísticos y su evolución lógica a lo largo de los últimos años y asumir una definición con la cual operar al hablar de corpus de aprendientes.

Como punto de partida, asumimos la propuesta de Rojo (2021), que incluye de forma concisa los rasgos más prototípicos de un corpus lingüístico. Según él, un corpus lingüístico 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 (Rojo, 2021, p. 1).

La definición de Rojo (2021) enuncia algunas características de los corpus lingüísticos que podríamos enriquecer con las propuestas de caracterización ofrecidas por Bowker y Pearson (2002), Villayandre (2008) y Parodi (2008) que también hemos revisado en Gou Jialing, Rodríguez Roque y Cuba (2023). No obstante, 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 consideramos contenidas en la definición de corpus lingüísticos ofrecida por Rojo (2021) y que podemos desglosar de la siguiente manera:

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

Defendemos que todos los corpus lingüísticos deberían respetar estos rasgos prototípicos definitorios de lo que se entiende por corpus en el campo de la lingüística de corpus moderna. No obstante, deberían incluir también otras características específicas congruentes con su objetivo concreto y diseño preestablecido, teniendo en cuenta la variedad de corpus existentes.

Esta variedad podemos constatarla en las diversas tipologías que desde diferentes perspectivas (Torruella y Llisterri, 1999; Villayandre, 2008; Cruz, 2012; Hincapié y Bernal, 2018 y Rojo, 2021) buscan clasificar los corpus existentes. En este trabajo, consideramos pertinente la tipología de corpus ofrecida por Rojo (2021), ya que parte de los diferentes objetivos para los que se conciben los corpus, su diseño y otra serie de propiedades (tipos de textos incluidos y sus características, número de lenguas, grado de información añadida, etc.), lo que consideramos de gran valor.

Rojo (2021) propone que los corpus pueden ser de referencia o especializados, orales, escritos o multimodales, totales o muestrales, cerrados o abiertos, monolingüe o multilingüe, sincrónicos o diacrónicos, codificados o no codificados y anotados y no anotados.

#### *Hacia una definición de corpus de aprendientes*

De acuerdo con la tipología de Rojo (2021), los corpus de aprendientes constituyen un tipo de corpus especializado que, como otros, responde a un diseño específico. Pero, tal diseño debe partir de una definición que precise sus rasgos prototípicos.

Ya en Gou Jialing, Rodríguez Roque y Cuba (2023) nos referíamos a la falta de estandarización respecto a la definición de corpus de aprendientes en la

bibliografía especializada. Por ello, partiendo de una revisión de las definiciones más citadas en la bibliografía referida a los corpus de aprendientes (Granger, 2002; Hincapié, 2018; Calero *et al*, 2020; Rojo, 2021 y Rojo y Palacios, 2022) y, sobre todo, tomando como base la definición general de corpus lingüístico (Rojo, 2021), en este trabajo consideramos que un corpus de aprendientes es un tipo de corpus especializado, constituido por un conjunto de textos, orales o escritos reales (y textos multimodales), 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.

De esta definición, cabe aclarar lo que entendemos por dos de sus rasgos constituyentes. Por un lado, cuando nos referimos a actos lingüísticos producidos en condiciones naturales lo entendemos en contraste con la práctica de inventar muestras de habla (Rojo y Palacios, 2022). Por tanto, estas condiciones naturales no necesariamente tienen que llegar a ser las que exige un corpus general del español que reúnen textos producidos en la lengua materna de los informantes. Por otro lado, al referirnos a que los textos deben ser producidos por estudiantes de una determinada lengua segunda (L2) o extranjera (LE) también hemos considerado la idea de que los informantes pueden ser también aprendientes de su lengua materna (Cruz, 2012), pero preferimos reservar el término corpus de aprendientes para los que recogen muestras de la interlengua en una lengua extranjera.

## Resultados

### *Generalidades del corpus CACHE*

El corpus CACHE fue diseñado con el objetivo de reunir textos en formato digital, representativos de las modalidades oral y escrita de ELE y EL2 de los aprendientes chinos y codificados con metadatos y anotados lingüísticamente,

para llevar a cabo investigaciones rigurosas sobre su interlengua y utilizarlo como recurso didáctico.

Una vez completado, el corpus CACHE va a constituirse como un corpus especializado, muestral y representativo de la interlengua de los estudiantes chinos de español. Será multilingüe (incluye un corpus nativo de control en español y chino mandarín), mixto (incorpora muestras orales y escrita), abierto y además será un corpus de orientación transversal.

Está conformado por cuatro subcorpus:

- Subcorpus de Aprendientes Chinos de Español como Lengua Extranjera (SUBCACHELE): participan informantes chinos que estudian español en un contexto sin inmersión lingüística.
- Subcorpus de Aprendientes Chinos de Español como Segunda Lengua (SUBCACHEL2): colaboran informantes chinos que aprenden español en contextos de inmersión.
- Subcorpus de Hablantes Nativos de Español (SUBCNat.E): recoge muestras de lengua de hablantes nativos de español.
- Subcorpus de Hablantes Nativos de Chino (SUBCNat.CH): recopila muestras de lengua de hablantes de chino mandarín.

277

Cada uno de ellos incluye un componente oral y otro escrito.

En este trabajo describimos CACHE de manera general, pero nos centraremos en la descripción de SUBCACHELE Escrito, que es el subcorpus más avanzado hasta la fecha.

#### *Descripción de sitio web del CACHE*

La construcción tecnológica de CACHE requirió la colaboración de la autora con un equipo de especialistas en informática y lingüística. Del trabajo de construcción tecnológica, a continuación, se destacan los elementos más importantes: el desarrollo *frontend* se llevó a cabo con *ReactJS* y *Next.js*, el almacenamiento de datos con *MongoDB*, el diseño visual con *HTML5* y *CSS*, la

implementación y despliegue en *Vercel* y el desarrollo del control de versiones con *Git*.

El sitio web del corpus CACHE fue creado utilizando *Vercel*. Su primera versión está disponible en línea (<https://corpus-nivald2leon.vercel.app/#>) para todos los investigadores y profesores interesados en el estudio y perfeccionamiento de la interlengua de los estudiantes chinos que aprenden ELE. Tras destacar esta característica de acceso libre, queremos describir brevemente el sitio web construido.

En la página principal de CACHE destaca una nube de palabras que recrea el mapa de China. En ella se agrupan y relacionan diferentes términos, entre ellos, las denominaciones del corpus, sus contenidos y sus aplicaciones para la investigación y didáctica de ELE.

La barra superior de la página principal contiene un menú general que, como muestra la Figura 1, incluye 8 elementos: Inicio, Concepto, Diseño, Estadísticas, Cache, Guía, Investigaciones y Contacto.

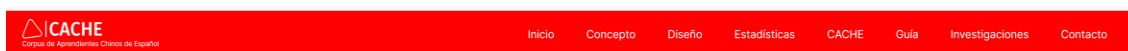


Fig. nº 1. Menú general del sitio web CACHE

“Inicio” redirige a la página principal de CACHE descrita con anterioridad. Por su parte, al dar clic en “Concepto” despliega un submenú que incluye otros dos elementos: *¿Qué es CACHE?* y *Objetivo*. Obsérvese la Figura 2.

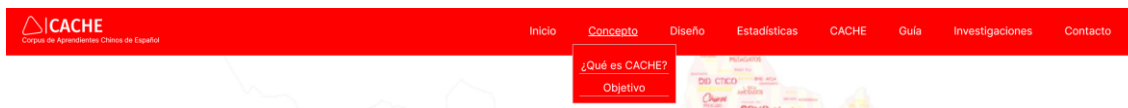


Fig. nº 2. Submenú de “Concepto” del sitio web CACHE

El primero de estos elementos redirige a una descripción general del corpus CACHE: el significado de la abreviatura, sus características y la definición de corpus de aprendientes de la que se parte para su concepción. Por otro lado, al dar clic sobre el segundo de estos elementos, se redirige a una página interna en la que aparece el objetivo general de CACHE.

En “Diseño”, como se exhibe en la Figura 3, se despliega un submenú con cinco elementos, a saber, *Diseño macro*, *Fundamentos teóricos y metodológicos*, *Etapas*, *Metadatos lingüísticos y extralingüísticos* y *Tareas*.



Fig. nº 3. Submenú de DISEÑO del sitio web CACHE

Cada uno de estos elementos del submenú, redirige a una página interna que incluye el contenido correspondiente a la denominación: en *Diseño macro* se presenta una descripción general del diseño de CACHE, en forma de texto y en esquema, que sirve de forma orientativa para conocer la estructura del corpus construido y su concepción desde su diseño; en *Fundamentos teóricos y metodológicos* aparece una página interna donde se exponen los principios que sirvieron de base para la construcción de CACHE; en *Etapas* se presenta la metodología de la construcción del corpus; en *Metadatos lingüísticos y extralingüísticos* se exponen todos los metadatos utilizados para la codificación de los textos recopilados en SubCACHELE Escrito que es hasta ahora el único subcorpus y componente construido; por último, en *Tareas* se presentan todas las actividades de redacción empleadas para la recopilación de los textos.

El cuarto elemento del menú general es “Estadística”. Este botón redirecciona a una página interna donde aparecen tabulados los datos estadísticos actuales de CACHE. Hasta el momento solo se describen las estadísticas del SubCACHELE Escrito, porque los restantes subcorpus y componentes están en construcción. La estadística descrita la actualizaremos periódicamente porque CACHE es un corpus abierto al que iremos incorporando datos a medida que los recopilamos y codifiquemos.

El quinto elemento del menú, denominado CACHE, incluye un submenú con 8 elementos, a saber, SUBCACHELE Escrito, SUBCACHELE Oral,

SUBCACHEL2 Escrito, SUBCACHEL2 Oral, SUBCNat.E Escrito, SUBCNat.E Oral, SUBCNat.CH Escrito y SUBCNat.CH Oral. Obsérvese la Figura 4.



Fig. nº 4. Submenú de CACHE del sitio web CACHE

Cada uno de estos elementos, cuyas denominaciones aluden a los diferentes subcorpus y los componentes que integran CACHE, redirecciona a la página interna de búsqueda en cada subcorpus. Así, por ejemplo, en el SubCACHELE Escrito aparece la siguiente página interna que se presenta en la Figura 5.

Fig. nº 5. Página de búsqueda del SubCACHELE Escrito sitio web CACHE

El buscador de SubCACHELE Escrito permite realizar búsquedas avanzadas a partir de filtros gracias a la codificación de los textos a partir de los metadatos lingüísticos y extralingüísticos establecidos desde el diseño de CACHE. Por ejemplo, es posible llevar a cabo una búsqueda de la construcción verbal «Creo que» seleccionando el nivel de competencia de los estudiantes, la tarea de redacción, el género textual predominante, así como filtrando mediante otras variables como el sexo, la edad y el nivel de instrucción de los informantes, etcétera. Seguidamente, en la Figura 6, ofrecemos un ejemplo de la búsqueda de «Creo que» a partir de metadatos específicos.

Subcorpus de Aprendientes Chinos de Español como Lengua Extranjera. Componente Escrito.

**Limpiar** **BUSCAR**

Nivel de competencia en español: B2 (Avanzado)  
 Tema de redacción: Opción 1. Usted ha leído un mensaje e...  
 Género textual: Expositivo y descriptivo  
 Sexo: Masculino  
 Nivel de institución: Preuniversitario  
 Dominio de lenguas extranjeras: Inglés

Palabras (siempre obligatorias): Creo que  
 Segunda lengua materna: Mandarín del suroriente  
 Edad: 21  
 Entusiasmo: 17 18 19 20 21 22 23 24 25 26 27 28  
 Motivación por el estudio del español: 8  
 Años estudiando español: 3  
 0 1 2 3 4 5 6 7 8 9 10 11

Fig. nº 6. Muestra de la búsqueda de «Creo que» en el SubCACHELE Escrito sitio web CACHE

Una vez seleccionados estos metadatos para efectuar la búsqueda de «Creo que», se da clic en el botón BUSCAR. Como se observa en la Figura 7, aparece una lista de ocurrencias de «Creo que».

Subcorpus de Aprendientes Chinos de Español como Lengua Extranjera. Componente Escrito.

111. Me gusta leer porque mis padres me enseñan desde pequeño; además, me encanta hacer deporte. **creo que** la vida está en movimiento. Es un placer conocerlos. [...]

113. ¿Judo español en China. Soy de Anhui. Hay tres personas en mi familia: mi padre, mi madre y yo. **Creo que** soy una chica amable y alegre. En mi tiempo libre, me gusta pintar, tocar la guitarra y hacer [...]

114. el pueblo que lo habla. Como siempre quería explorar lo desconocido y cada lengua es un mundo, **creo que** lo mejor formar de explorar el mundo es aprender las lenguas con las que las personas interactúan. [...]

116. ¿Cuál es el significado de familia? **Creo que** es un destino. La familia es un regalo especial que nos distribuye los dioses. Y cada una es de [...]

119. **Creo que** el mejor profesor es Diabul, nos enseña a español, ha estado con nos alumnos desde que entramos. [...]

1107. ¡Hola, soy Dorian. ¿Y tú? Me gusta leer la literatura china, sobre todo, sobre de Wang Zengqi. **Creo que** sus cuentos y ensayos son muy precisos. También leo muchas novelas de Maugham. Me encanta ver [...]

1125. ¡Javier! ¿pero últimamente me ocupas en el examen final. Ya que a papá le ha encantado Madrid, **creo que** pueden pasar las vacaciones allí. En la ciudad hay famosos museos y también pueden tomar espes [...]

1132. ¡EEE-4 pero los exámenes finales van a empezar la próxima semana. Sin embargo, no te preocupes. **Creo que** puedo hacerlo bien. Además, estoy muy alegre de que tengáis una vacación tan larga. Me gusta! [...]

1134. ¡men muy importante. Me alegro de saber que vosotros van a viajar y realmente quiero ir contigo. **Creo que** podemos ir a Valencia, la tercera ciudad más grande de España, conocida como la Ciudad del Sol [...]

1139. ¡otro divertido. Tengo muchas ganas de ir allí para ver la Ciudad Prohibida y la Gran Muralla. **Creo que** también te interesará la historia y la cultura china. Además, Miami también es una buena opción. [...]

1146. ¡ HELTS, pero he aprendido mucho y me siento muy realizado. En cuanto a irnos de viaje juntos, **creo que** estaría más que encantado de irme de viaje contigo cuando acabe mis exámenes, ya que me gusta! [...]

1150. ¡buenos de nada en tu apartamento actual. Mi piso tiene espacio suficiente para dos personas y **creo que** podremos llevarnos bien completamente juntos. Sin embargo, odio ser torcido y decirte que no! [...]

1160. ¡bueno abandonar alquilar el piso y hay una habitación vacía afuera. No es un piso muy grande pero **creo que** es suficiente para ti y la perro. Ah una cosa, me gusta pero mucho, por eso no te preocupes. [...]

1172. ¡ competición política el martes. Ayer hice un examen. Estoy tan cansado y necesito un descanso. **Creo que** será libre la próxima semana. Espero que podamos ir al restaurante español juntos el próximo m [...]

1176. ¡bueno, la semana que viene podemos quedar porque todavía me quedan muchos días para el examen. **Creo que** el próximo viernes podemos quedar y también espero mucho verte. ¡hasta la próxima semana! [...]

1179. ¡ Pero, ¡qué suerte! La próxima semana terminará el curso. Entonces, tendrá mucho tiempo libre. **Creo que** es más conveniente ir viernes o el sábado. ¿Está bien el viernes para ti? ¿adónde vamos? [...]

1182. ¡ayudar a mi maestra a hacer experimentos. [...]

Fig. nº 7. Muestra de las ocurrencias de «Creo que» en el SubCACHELE Escrito sitio web CACHE

Cuando se seleccionada una de estas ocurrencias de «Creo que», se redirige a una página interna donde aparece el texto completo con el elemento lingüístico objeto de búsqueda resaltado en negrita y todos los metadatos del texto. Sirva como muestra la Figura 8.



Fig. nº 8. Muestra de un texto en el que están dos de las ocurrencias de «Creo que» en el SubCACHELE Escrito sitio web CACHE

El sexto elemento del menú principal de CACHE es la GUÍA. Al dar clic sobre este elemento se redirecciona a una página interna en la que se orienta a los usuarios del corpus de la manera en que se realizan las búsquedas en el buscador avanzado. Obsérvese la Figura 9.



Fig. nº 9. Muestra de la vista de la GUÍA del SubCACHELE Escrito sitio web CACHE

El antepenúltimo elemento del menú general de CACHE es “Investigaciones”, que es una página interna donde se recogen las investigaciones o unidades didácticas realizadas con CACHE. Hasta ahora se han añadido dos publicaciones (Gou Jialing, 2022; Gou Jialing, Rodríguez Roque y Cuba, 2023) y una investigación en curso.



Fig. nº 10. Muestra de la vista de “Investigaciones” del sitio web CACHE

Por último, como se observa en la Figura 11, en “Contacto” se ofrece un formulario en línea que facilita la comunicación. Los investigadores y profesores interesados en usar o colaborar con el corpus podrían ponerse en contacto con nosotros a través de un buzón de correo electrónico.



Fig. nº 11. Muestra de la vista de “Contacto” del sitio web CACHE

### *Estadísticas generales de SubCACHELE Escrito*

SubCACHELE Escrito consta de 567 textos, hasta marzo de 2024. Como se observa en la Figura 12 (Gráfico 1), en total recopilamos 83.655 palabras. La meta de este subcorpus es alcanzar como mínimo el millón de palabras, pero para ello necesitamos la colaboración de equipos de trabajo de diferentes universidades de China donde se estudia la carrera de Filología Hispánica y hasta ahora solo se ha trabajado con seis universidades.

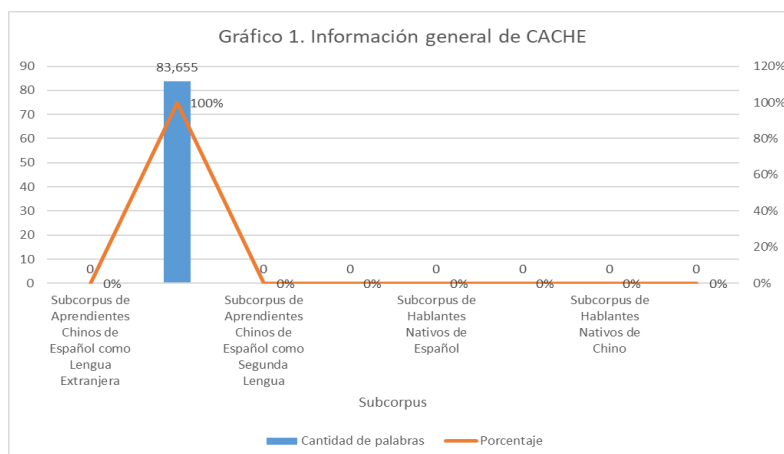


Fig. nº 12. Gráfico 1. Información general de CACHE

Como se observa en la Figura 13 (Gráficos 2A y 2B), de acuerdo con el nivel de competencia, codificamos 119 textos (20.99%) del nivel A1 (9.388 palabras), 104 (18.34%) del A2 (9.516 palabras), 175 (30.86%) del B1 (27.559 palabras), 115 (20.28%) del B2 (20.156 palabras), 34 (6.00%) del C1 (10.022 palabras) y 20 textos (3.53%) del nivel C2 (7.014 palabras).



Fig. nº 13. Gráfico 2A y 2B. Textos y palabras según el nivel de lengua

Cabe señalar que la mayor cantidad de palabras se distribuyen entre los niveles B1 y B2, con 27.559 y 20.156 palabras, respectivamente, lo que significa un 32.94% y un 24,09% del total (83.655 palabras, 100%). Los niveles de competencia de los que se lograron recopilar menos datos con el C1 y C2, de los que se han recogido solamente 10.022 y 7.014 palabras, correspondientemente, lo que representa un 11.98% y un 8.38% del total (83.655 palabras, 100%).

### *Estadísticas según los metadatos (extra)lingüísticos de SubCACHELE Escrito*

Siguiendo con la descripción estadística del SubCACHELE Escrito, en la Figura 14 (Gráficos 3A y 3B) se observa la cantidad de textos según el género textual predominante y la distribución de palabra para cada caso. De los 567 codificados, hay 374 (65,96%) textos expositivos-descriptivos para un total de 46.675 (55,79%) palabras, 68 (11,99%) textos narrativos donde se recogen 9.654 (11,54%) palabras y 125 (22,05%) textos argumentativos que incluyen 27.326 (32,67%) palabras. Por tanto, en más de la mitad de los textos recopilados en SubCACHELE Escrito predomina el género expositivo-descriptivo.

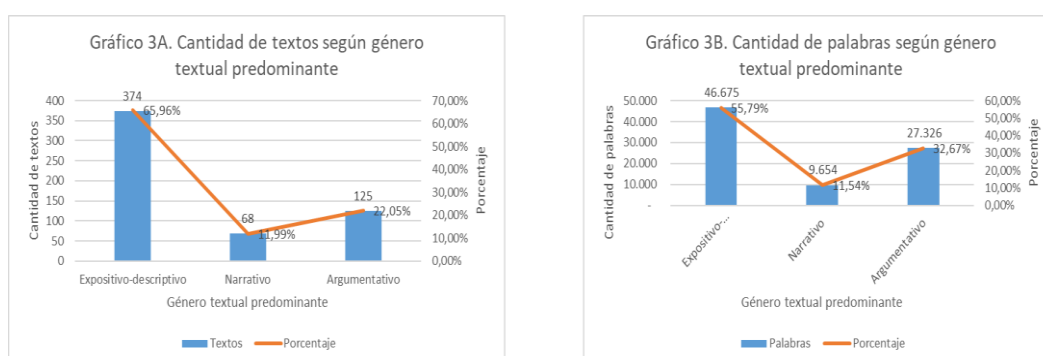


Fig. nº 14. Gráfico 3A y 3B. Textos y palabras según el género textual predominante

Respecto a las variables tiempo de estudio de ELE y estancia en países hispanos, puede consultarse en la Figura 15 (Gráficos 4A y 4B). De los datos tabulados, pueden destacarse que el tiempo de estudio de los participantes oscila entre uno y ocho años. La mayoría de los textos fueron escritos por estudiantes que llevan de uno a cuatro años estudiando ELE. Esto queda explicado en que pertenecen a la matrícula activa de las universidades involucradas en el momento de recolección. Cabe resaltar que los estudiantes que llevan más de 5 años estudiando ELE son graduados de estas mismas universidades.

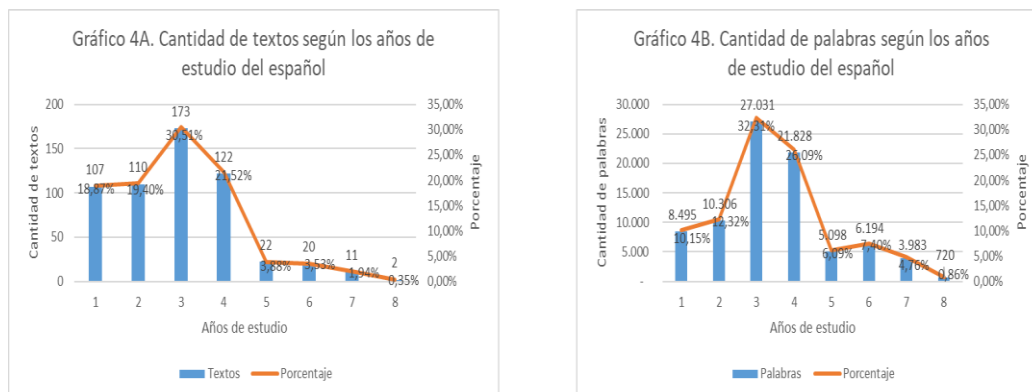


Fig. nº 15. Gráfico 4A y 4B. Textos y palabras según los años de estudio del español

Por otra parte, la estadística referida a la estancia de los informantes en países hispanos queda resumida en la Figura 16 (Gráficos 5A y 5B). La mayoría de los estudiantes no han hecho estancias de estudio ni intercambios. Del total de los textos recopilados, 473 (83%) fueron escritos por estudiantes que nunca han estudiado fuera de China, 67 textos (12%) por estudiantes que han estado en España por 1 año, 18 textos (3%) y 9 textos (2%) fueron producidos por estudiantes con 2 y 3 años de estancia, respectivamente.

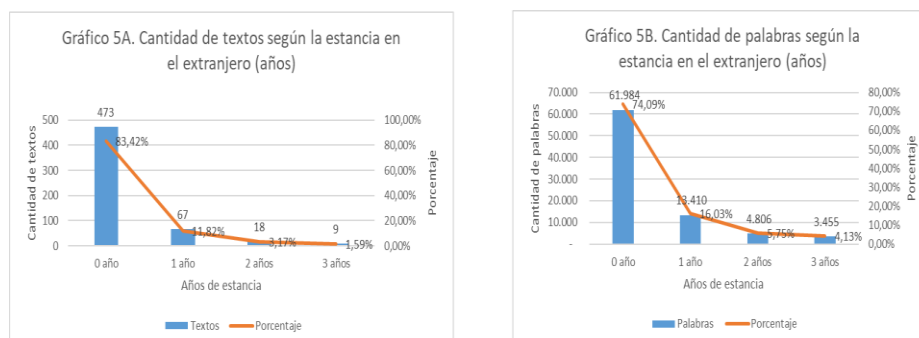


Fig. nº 16. Gráfico 5A y 5B. Textos y palabras según la estancia en el extranjero

En relación con las variables sociales sexo, edad y nivel de instrucción, sobresalen los siguientes valores:

Como se observa en la Figura 17 (Gráficos 6A y 6B), respecto a la variable sexo, se recopilaron más textos producidos por mujeres que por hombres, lo que resulta coherente con el hecho de que en la carrera de Filología Hispánica el porcentaje de mujeres es mayor que el de hombres. De los 567 (100%)

textos recopilados, 335 (59.08%) fueron producidos por mujeres y 232 textos (40.92%) por hombres. De estos datos, en cantidad de palabras, la distribución sería: 49.246 palabras (58.87%) en los textos escritos por las mujeres y 34.409 palabras (41.13%) en los de los hombres.



Fig. nº 17. Gráfico 6A y 6B. Textos y palabras según el sexo

En lo concerniente a la variable edad, como se observa en la Figura 18 (Gráficos 7A y 7B), la edad de los participantes oscila entre 18 y 26 años, pero la mayoría de los textos, 474 (83.59%) fueron escritos por estudiantes de 19 a 22 años, que se encontraban estudiando la carrera de Filología Hispánica en el momento de recogida de datos.

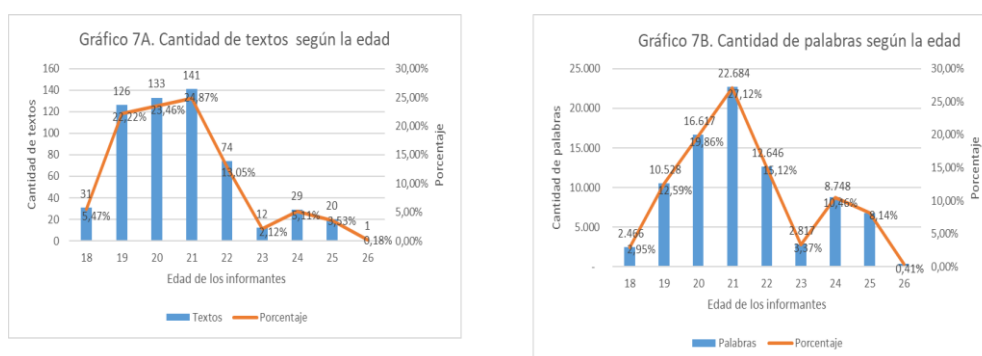


Fig. nº 18. Gráfico 7A y 7B. Textos y palabras según la edad

En cuanto al nivel de instrucción, registrado en la Figura 19 (Gráficos 8A y 8B), de los 567 textos (100%) recopilados, 501 (88.36%) pertenecen a estudiantes graduados del bachillerato que cursaban la carrera de Filología Hispánica en las universidades involucradas en este estudio, 59 textos (10.41%) a estudiantes graduados de Filología Hispánica por esas mismas universidades y

que se encontraban estudiando el Máster en Filología Hispánica en China, 7 textos (1.23%) a estudiantes con nivel de instrucción de máster, graduados de la carrera de Filología Hispánica en estas universidades y el Máster en Filología Hispánica en China y que actualmente se encuentran en programas de doctorados en especialidades relacionadas con ELE. De estos datos, pero en términos de número de palabras, cabe enfatizar que 64.292 (76.85%) de las palabras recopiladas pertenecen a informantes del nivel de instrucción preuniversitario.

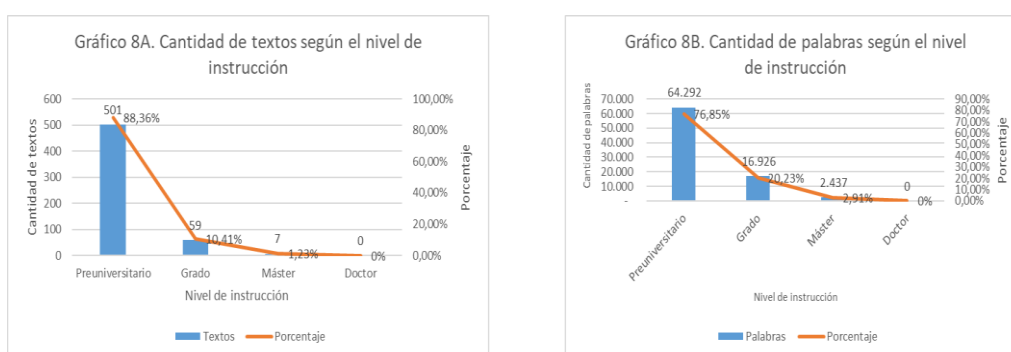


Fig. nº 19. Gráfico 8A y 8B. Textos y palabras según el nivel de instrucción

## Discusión

En este trabajo, redefinimos el término corpus de aprendientes, retomamos las etapas establecidas para la creación de CACHE y, desde el punto de vista práctico, describimos tecnológicamente y estadísticamente el SubCACHELE Escrito. Hasta marzo de 2024, el subcorpus consta de 567 textos (83.655 palabras, 100%). Cada texto fue codificado a partir de distintas variables (extra)lingüísticas. Cabe destacar que, de acuerdo con el nivel de competencia, se codificaron 119 textos (20,99%) del nivel A1, 104 (18.34%) del A2, 175 (30.86%) del B1, 115 (20.28%) del B2, 34 (6.00%) del C1 y 20 textos (3.53%) del nivel C2. Esta estadística permite tomar decisiones sobre posibles investigaciones y aplicaciones didácticas que podríamos llevar a cabo, teniendo en cuenta el nivel de los informantes. También, nos indican el camino que debemos seguir para ampliar la muestra de la interlengua de los estudiantes chinos de Filología Hispánica.

En cuanto a las aplicaciones del SubCACHELE Escrito, su uso en el estudio de la interlengua de los estuantes chinos y en la enseñanza-aprendizaje de ELE en el contexto universitario chino puede ser muy variado. Con este subcorpus de aprendientes, como plantean Rojo y Palacios (2022), podemos perseguir siempre un doble objetivo: por un lado, podemos intentar conocer mejor los rasgos que presentan las interlenguas propias de cada uno de los grupos incluidos en el subcorpus; por otro, podemos utilizar este conocimiento para identificar mejor las zonas conflictivas y diseñar entonces estrategias más adecuadas para su mejor solución.

Es cierto que uno de los usos directos de este tipo de corpus de aprendientes suele ser la investigación del estado de la interlengua. Pero, en este trabajo queremos insistir en que sus aplicaciones pueden ir mucho más allá. De hecho, cada vez son más frecuentes posicionamientos que defienden su utilidad como recursos didácticos en la elaboración de materiales didácticos, el diseño curricular de cursos, la formación de profesores e incluso la confección de exámenes. En este sentido, queremos destacar las aplicaciones propuestas por Rojo y Palacios (2022):

- Los datos extraídos de estos corpus pueden facilitar información complementaria de interés para la creación de manuales, gramáticas, diccionarios y medios didácticos para el aprendizaje del español, por ejemplo, glosarios elaborados de forma específica para estudiantes con determinado perfil.
- Los datos proporcionados por el corpus de aprendientes poder servir para alertar sobre posibles falsos amigos, voces con ortografía irregular y que resultan complejas para los estudiantes, además, problemas comunes en el uso de determinadas cuestiones gramaticales como el uso de preposiciones como *para* y *por*, las diferencias entre *ser* y *estar*, e incluso la utilización de distintos términos en adecuación con un registro formal o informal, etc.
- La formación y superación del profesorado también se pueden organizar partiendo de muestras extraídas de los corpus de aprendientes, por ejemplo, hay que enseñar a clasificar a los estudiantes de acuerdo con los niveles de lengua establecidos en los descriptores del MCER (2002, 2021) o hay que formar a los profesores en la concepción de diversas

formas de tratamiento didáctico ante determinadas dificultades en el aprendizaje, lo que puede ilustrarse con muestras concretas seleccionadas del corpus.

- Por último, pueden mencionarse al menos otras tres potencialidades de los corpus de aprendientes: el diseño curricular de programas, cursos y carreras, la elaboración de exámenes y el diseño de herramientas computacionales de detección de errores. No cabe duda de que estas acciones se beneficiarían de los datos que proporcionan los corpus de aprendientes.

Estas son, por supuesto, solo aplicaciones generales de los corpus de aprendientes de las que se pueden desprender variadísimas propuestas que demuestran la importancia de la creación del SubCACHELE Escrito como fuente de datos empíricos para la investigación de la interlengua de los estudiantes de Filología Hispánica y como recurso didáctico para el perfeccionamiento de la enseñanza universitaria del español en China.

## Referencias Bibliográficas

- Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4), 243-257. <https://doi.org/10.1093/lc/8.4.243>
- Biber, D., Conrad, S., y 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., y Pearson, J. (2002). *Working with Specialized Language A Practical Guide to Using Corpora* (1era ed.). Routledge.
- Calero F., M A.; Serrano Z., M<sup>a</sup> I. y Gómez-Devís, M 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](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/doblele.35>
- Francis, W. N., Kucera, H., y Mackie, A. W. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin. <https://lib.ugent.be/en/catalog/rug01:000049253>
- Gou, J. (2023). Diseño de un corpus de aprendientes chinos de español: potencialidades para el estudio y el tratamiento de la atenuación en español. [Tesis de doctorado en Ciencias Lingüística, Universidad de La Habana]. Repertorio de la Universidad de La Habana.
- Gou, J., Rodríguez Roque, D., y Cuba V., 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>
- Granger, S. (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. John Benjamins Publishing Company. <https://doi.org/10.1075/llt.6>
- Gries, S. T., y Stefanowitsch, A. (2006). *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Mouton de Gruyter. <https://n9.cl/jv7cj>
- Gries, S. (2009). What Is Corpus Linguistics. *Language and Linguistics Compass*, 3, 1–17. <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.
- 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://n9.cl/owio7>
- Leech, G. (1991). The state of the art in corpus linguistics. En K. Aijmer y B. Altenberg (Eds.). *English Corpus Linguistics. Studies in honor of Jan Svartvik* (pp. 8-29). Longman.
- 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. 155-170). John Benjamins Publishing Company.
- Lemnitzer, L., y 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, T. (2001). *Corpus Linguistics: An Introduction*. Edinburgh University Press.
- McEnery, A. M., Xiao, R. Z. and Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge.
- Parodi, G. (2008). Lingüística de corpus: Una introducción al ámbito. *Revista de lingüística teórica y aplicada*, 46(1), 93-119.  
<https://doi.org/10.4067/S0718-48832008000100006>
- Rojo, G. (2021). *Introducción a la lingüística de corpus en español*. Routledge.
- 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). Routledge.
- Santalla del Río, M. Paula. (2005). La elaboración de corpus lingüísticos. En: Cal, M., Núñez, P., y Palacios, I. *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de Lenguas* (pp. 45-62). Servizo de Publicacións da Universidade de Santiago de Compostela.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. (1996). *Preliminary recommendations on Corpus Typology*. EAGLES.
- Sinclair, J. (2005). Corpus and Text: Basic Principles. En M. Wynne (ed). *Developing Linguistic Corpora: A guide to good practice* (pp. 1-16). Oxbow Books.
- 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.* (pp. 45-77).  
Universidad Autónoma de Barcelona. Milenio.  
Villayandre Ll., M. (2008). Lingüística con corpus (I). *Estudios humanísticos. Filología*, 30, 329-349. <https://doi.org/10.18002/ehf.v0i30.2847>