



RAEL: Revista Electrónica de Lingüística Aplicada

Vol./Núm.: 24/1
Enero-diciembre 2025
Páginas: 132-150
Artículo recibido: 17/09/2024
Artículo aceptado: 20/05/2025
Artículo publicado: 31/12/2025
Url: <https://rael.aesla.org.es/index.php/RAEL/article/view/620>
DOI: <https://doi.org/10.58859/rael.v24i1.620>

Aplicaciones de los corpus en la enseñanza del chino como lengua extranjera: un estudio exploratorio

Applications of Corpora in the Teaching of Chinese as a Foreign Language: An Exploratory Study

LINGZHI NIE

SARA ROVIRA-ESTEVA

UNIVERSIDAD AUTÓNOMA DE BARCELONA

Aunque el interés por el chino ha experimentado un gran auge en los países hispanohablantes en las últimas décadas, este aumento en el número de estudiantes no ha venido acompañado de un incremento significativo de la investigación, en general, y en lingüística de corpus, en particular, que acabe transfiriéndose a la sociedad en forma de innovación docente. Así pues, para llenar este vacío y ante las necesidades y nuevos retos que enfrentamos para la enseñanza del chino como lengua extranjera, el objetivo de este estudio es revisar la oferta de corpus de chino existentes y, a la vez, ofrecer algunos ejemplos prácticos de cómo se pueden aplicar este tipo de recursos a la enseñanza de chino como lengua extranjera para sacarles el máximo partido y abrir nuevas perspectivas tanto en su investigación como en su docencia.

Palabras clave: *lingüística de corpus; aplicaciones de corpus; corpus de chino; chino como lengua extranjera; CLE.*

Even though interest in Chinese has boomed in Spanish-speaking countries in the last few decades, this increase in the number of students has not been accompanied by a significant increase in research in general, and in corpus linguistics in particular, which would be transferred to society in the form of teaching innovation. Therefore, to fill this gap and in view of the needs and new challenges educators face in teaching Chinese as a foreign language, the aim of this study is to review the existing Chinese corpora and, at the same time, to provide some practical examples of how these types of resources can be applied to the teaching of Chinese as a foreign language to make the most out of them and to open up new perspectives for both research and teaching.

Keywords: *corpus linguistics; applications of corpora; corpora of Chinese; Chinese as a foreign language; TCFL.*

Citar como: Nie, L., & Rovira-Esteva, S. (2025). Aplicaciones de los corpus en la enseñanza del chino como lengua extranjera: un estudio exploratorio. *RAEL: Revista Electrónica de Lingüística Aplicada*, 24, 132-150. <https://doi.org/10.58859/rael.v24i1.620>

1. INTRODUCCIÓN

De acuerdo con McEnery et al. (2010: 4), un corpus es “[a] collection of sampled texts, written or spoken, in machine-readable form which may be annotated with various forms of linguistic information”. Es decir, además de ser una colección de textos ordenados de una determinada manera, ya sean científicos, literarios, entre otros, son analizables por las máquinas y pueden estar anotados con información metalingüística para satisfacer necesidades académicas.

Fundamentalmente, a diferencia de la gramática generativa transformacional de Chomsky del *what is possible*, la lingüística de corpus parte más de la idea del *what is probable* (Kennedy, 1998: 7-8), argumentando que las oraciones que no se corresponden con el uso real tienen una baja probabilidad de aparecer, aunque sean gramaticalmente correctas.

Las aplicaciones de los corpus son mucho más amplias de lo que puede parecer a primera vista. De acuerdo con Biber et al. (1998: 11), la investigación con corpus puede beneficiar a casi todas las ramas de la lingüística. Hunston (2002: 96) afirmó que incluso personas que nunca han oído hablar de corpus están usando productos derivados de la investigación basada en corpus. El ejemplo más ilustrativo son los diccionarios, que pueden ser elaborados a partir de corpus, sobre todo los ejemplos de uso de palabras y expresiones. A parte de la lexicografía, los corpus también se utilizan en la lingüística contrastiva, la traductología, la traducción automática, la enseñanza y el aprendizaje de idiomas, así como el entrenamiento de *chatbots*, entre otros.

Existen numerosos estudios sobre el uso pedagógico de los corpus. Por ejemplo, Conrad (2000) exploró el impacto de la lingüística de corpus en la docencia de la gramática; Belz y Vyatkina (2005) presentaron una intervención pedagógica basada en corpus para la enseñanza de las partículas modales del alemán; Farr (2008) valoró el uso de corpus electrónicos en la enseñanza del inglés en función de la experiencia de los usuarios; Boulton y Cobb (2017) investigaron la eficacia del uso de las herramientas y técnicas de corpus para el aprendizaje de segundas lenguas mediante un estudio meta-analítico; y, por su parte, Zhou et al. (2017) analizaron la construcción y aplicación de los corpus para la creación de materiales didácticos de chino como lengua extranjera (CLE).

Al disponer de una gran variedad de corpus que cubren un gran abanico de áreas de especialidad, estos pueden constituir una base de datos muy útil, tanto para la investigación como para la docencia del CLE, sea para la enseñanza de la lengua, en general, como de la lengua para fines específicos, en particular. Además, los corpus no solo nos proporcionan datos sobre el uso real del lenguaje, sino que también pueden ayudar a reducir el tiempo y la carga de trabajo del profesorado en la preparación de las clases. Sin embargo, a pesar del elevado número de corpus de lengua china y de publicaciones que tratan de sus aplicaciones para la enseñanza del CLE, estas contribuciones consisten mayoritariamente en estudios de caso, más que ofrecer visiones panorámicas. Prueba de ello es que la búsqueda llevada a cabo el 28 de febrero de 2024 con la palabra clave *corpus based* combinada con *Chinese teaching* en *Google Scholar* y sus equivalentes en chino *yuliaoku* (语料库) y *Hanyu jiaoxue* (汉语教学) en CNKI, la mayor base de datos académica china, arrojó un volumen considerable de resultados (690 y 3.635, respectivamente). En cambio, las palabras clave *corpus-based Chinese teaching*, *corpus for Chinese teaching* y su equivalente en chino *yuliaoku de Hanyu jiaoxue yingyong* (语料库的汉语教学应用) obtuvieron unos resultados muy limitados (0, 2 y 46, respectivamente), tanto en número como en relevancia para este trabajo de revisión.

Para llenar este vacío investigador y ante las nuevas necesidades y retos con los que nos enfrentamos en la enseñanza del CLE, en este estudio nos hemos planteado las siguientes dos preguntas de investigación: en primer lugar, cuál es la oferta actual de corpus de lengua china y, en segundo lugar, cómo se podrían aplicar estos corpus a la mejora de la enseñanza del CLE.

Para responder a estas dos preguntas, nos hemos planteado dos objetivos. Por una parte, hacer una revisión del estado de la cuestión para tener una visión panorámica de la oferta de corpus de chino existentes y, por la otra, explorar sus posibles aplicaciones al aula del CLE.

2. MARCO TEÓRICO-METODOLÓGICO

Leech (1997: 5-6) señaló que existen tres aplicaciones didácticas de los corpus: uso directo, uso indirecto y elaboración de corpus orientados a la enseñanza. Basándose en este autor, McEnery y Xiao (2011) realizaron una revisión bibliográfica profundizando en la aplicación de los corpus en la enseñanza de idiomas, especialmente del inglés. Estos autores destacaron la importancia de tener en cuenta datos empíricos y concluyeron que las aplicaciones indirectas de los corpus desempeñan un papel más importante que las directas, al dirigirse estas últimas a un alumnado más avanzado.

Por otra parte, existen varias propuestas de clasificación de los tipos de corpus. Puesto que nuestro objeto de estudio es el CLE, en esta investigación adoptaremos la propuesta de Liang et al. (2010: 4-5), que es la más extendida en China. De acuerdo con estos autores, los corpus se pueden dividir en *generales* o *temáticos*, según su contenido; en *sincrónicos* o *diacrónicos*, en función del enfoque temporal; en *escritos* u *orales*, atendiendo al canal; de *hablantes nativos* o de *estudiantes*, en función de la fuente de las muestras lingüísticas; en *monolingües*, *bilingües* (o *paralelos*) y *multilingües*, según el número de lenguas. Asimismo, hemos completado esta clasificación tomando de Villayandre (2008: 345-346) las categorías de corpus *cerrados* (estáticos) o *abiertos* (dinámicos),¹ en función de si las colecciones de textos permanecen fijas o se van actualizando, así como de corpus *simples* y *codificados*, según el procesamiento implicado. El análisis de los corpus se ha realizado de acuerdo con estas etiquetas, algunas de las cuales se han expandido para adaptarlas a la casuística de nuestra muestra.

El proceso para alcanzar los objetivos marcados ha consistido en diversas fases, algunas de las cuales se han desarrollado en paralelo. En primer lugar, se ha llevado a cabo una revisión bibliográfica para conocer el estado de la cuestión sobre la aplicación de los corpus a la enseñanza de lenguas extranjeras, en general, y del chino, en particular.

En segundo lugar, hemos realizado una búsqueda de los corpus de chino existentes. Tras una extensa revisión, hemos seleccionado 56 corpus creados por instituciones de referencia, con amplia difusión y adecuados para la enseñanza de la lengua china, es decir, que pueden emplearse para estudiar el uso real de chino por nativos o por estudiantes en diferentes contextos o registros. La mayoría de estos 56 corpus son de acceso abierto directo o previo registro, lo cual facilita su consulta y los hace más atractivos para los usuarios. Sin embargo, también hemos considerado pertinente incluir en nuestro análisis una serie de corpus de acceso limitado o comerciales por su calidad y porque cuentan con una cantidad considerable de artículos científicos derivados. La Figura 1 ilustra las características principales de los corpus de nuestra muestra. La base de datos con el análisis detallado de los 56 corpus de nuestra muestra puede consultarse en acceso abierto (véase Nie y Rovira-Esteva, 2025).²

¹ Se ha optado por los términos sinónimos de *estáticos* y *dinámicos* para evitar la confusión terminológica con el tipo de acceso.

² Cabe destacar que esta muestra no incluye corpus de chino creados en países hispanohablantes por gozar de menor difusión y presentar un acceso limitado. No obstante, por su posible interés para los lectores hispanohablantes, tres de ellos se mencionan en las conclusiones y se pueden encontrar en el Apéndice 1.

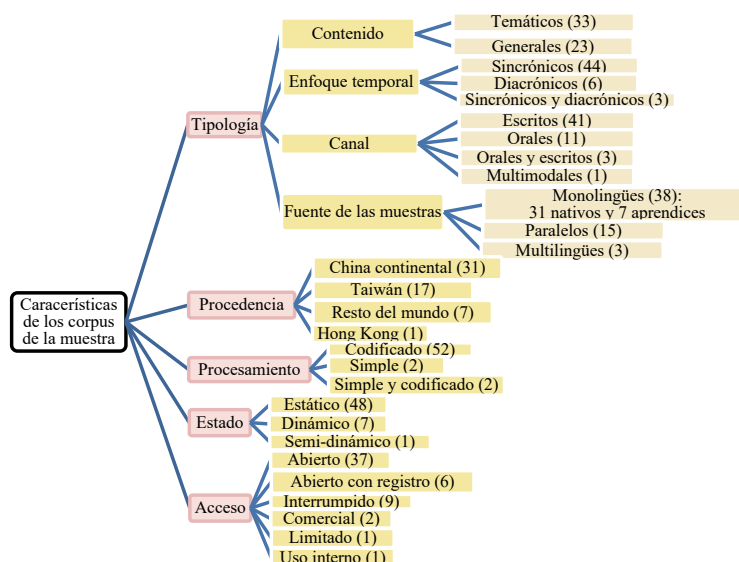


Figura 1: *Características principales de los corpus de la muestra*

Fuente: Elaboración propia a partir de la adaptación de Liang et al. (2010) y Villayandre (2008)

Finalmente, partiendo de los trabajos de Leech (1997) y McEnery y Xiao (2011) arriba mencionados y apoyándonos en la revisión bibliográfica sobre las aplicaciones de los corpus a la enseñanza del chino, así como los manuales de los corpus seleccionados, en este estudio presentamos una amplia gama de aplicaciones prácticas de los corpus existentes a la enseñanza del CLE, ilustrándolas con ejemplos concretos.

3. USOS DE LOS CORPUS EN LA ENSEÑANZA DEL CHINO COMO LENGUA EXTRANJERA³

3.1. Aplicaciones didácticas directas de los corpus

Leech (1997) explica las tres posibles aplicaciones directas de los corpus en los programas de lenguas y lingüística: 1) la enseñanza de la lingüística de corpus propiamente dicha (*teaching about*), 2) la formación en el uso de corpus (*teaching to exploit*) y 3) el uso de corpus como herramienta didáctica en el aula (*exploiting to teach*).

En cuanto a la primera, la lingüística de corpus, como metodología de análisis dentro de la lingüística, puede formar parte de los programas de estudio de lenguas extranjeras en niveles avanzados. Partiendo de los conocimientos y competencias necesarios para la investigación con programas de análisis de corpus, como *AntConc*, *WordSmith*⁴ y otros recursos y herramientas similares y la familiarización con la terminología relacionada (como *token*, *ocurrencia*, *KWIC* [*key words in context*], que indican un elemento de análisis, *treebank* para especificar el tipo de corpus o *frecuencia acumulativa* que es una aproximación estadística, entre otros).

La segunda aplicación directa consiste en formar a los estudiantes en métodos de aprendizaje enseñándoles a explotar los corpus para analizar o investigar el lenguaje. La tercera, en cambio, se centra en la optimización de los contenidos por parte del docente, aprovechando

³ La lista de todos los corpus mencionados en esta sección se puede encontrar en el Apéndice 1.

⁴ Véase <https://www.laurenceanthony.net/software/antconc/> y <https://wordsmith.org/>, respectivamente.

los corpus como recurso didáctico en la enseñanza de lenguas en el aula. Ambas aplicaciones están estrechamente relacionadas e incluso pueden ser complementarias. Por ejemplo, en el ámbito de la sociolingüística, en lugar de introducir aisladamente definiciones con ejemplos de extranjerismos, podemos recurrir a corpus para localizar ejemplos reales para la enseñanza o guiar a los alumnos en el uso de corpus para analizar sus características cualitativa y cuantitativamente. Asimismo, el análisis del discurso suele basarse en datos de corpus, por lo que es necesario elegir un corpus adecuado o elaborar un corpus propio conforme al objeto de análisis específico. En esta línea, Nie (en prensa) comparó los anglicismos en los medios de Hong Kong y Macao relacionados con el campo del entretenimiento a partir de un corpus de elaboración propia, mientras que Gong (2019) realizó un estudio contrastivo entre el chino y el español de las nominalizaciones en el discurso político.

El uso independiente de los corpus por parte de los estudiantes para descubrir patrones lingüísticos en el proceso de aprendizaje, es decir, el aprendizaje guiado por datos (*data-driven learning* [DDL]) y la práctica de los profesores de seleccionar, analizar y preparar materiales didácticos basados en la información de los corpus sin duda tienen similitudes con las últimas dos aplicaciones directas de los corpus, la primera incluso implica un contacto directo de los estudiantes con los corpus, mientras que la segunda refleja el papel de los recursos de los corpus como apoyo al contenido de la enseñanza y el aprendizaje. Sin embargo, ninguno de los dos enfoques consiste esencialmente en la introducción explícita de los corpus por parte del profesor como recurso didáctico en el aula. En concreto, el DDL tiene una función pedagógica que se puede considerar una extensión del aula, mientras que el uso de corpus para preparar materiales pertenece al trabajo previo al proceso de enseñanza propiamente dicho, en el que los alumnos no tienen contacto directo con los corpus. Por lo tanto, basándonos en el criterio de si los corpus se usan o no explícitamente en el proceso de enseñanza, este artículo clasifica estas dos situaciones como aplicaciones indirectas de los corpus en la enseñanza de idiomas, por lo que aparecen detalladas en el apartado siguiente.

3.2. Aplicaciones didácticas indirectas de los corpus

Siguiendo el esquema de Leech (1997), la principal aplicación didáctica indirecta de los corpus es la confección de diccionarios, que constituyen una herramienta esencial para la enseñanza y el aprendizaje de lenguas extranjeras. Sin embargo, también abarca el diseño de planes de estudios y la elaboración de materiales, el aprendizaje basado en datos, la evaluación de la competencia lingüística y la formación del profesorado. A continuación, explicaremos cada una de estas aplicaciones por separado, ilustradas con ejemplos de los corpus chinos de nuestra muestra.

3.2.1. Confección de diccionarios

McEnery y Xiao (2011: 367) afirman que el uso de corpus ha revolucionado la lexicografía y que el uso de ejemplos auténticos en los diccionarios facilita mucho el aprendizaje de palabras por parte de los estudiantes de idiomas extranjeros.

En la actualidad, es necesario profundizar en la aplicación de los corpus en la lexicografía china, ya que cuenta con un número limitado de diccionarios chinos elaborados a partir de corpus. El ejemplo más conocido es el uso del *NJU_BDRCBC corpus* (6),⁵ un corpus paralelo bilingüe inglés-chino desarrollado por el Bilingual Dictionary Research Center de la Nanjing University, que ha dado lugar a la colaboración con varias editoriales de obras lexicográficas

⁵ Para facilitar la búsqueda de los corpus citados, hemos indicado entre paréntesis el número el número que le corresponde en el Apéndice 1, donde están ordenados por el orden alfabético del autor o institución responsable.

de referencia. *El New Age English-Chinese Dictionary* (Pan, 2013), elaborado conjuntamente por este centro y la editorial Commercial Press, es el más famoso por ser el primer diccionario inglés-chino elaborado a gran escala a partir de un corpus.

Los tres primeros diccionarios monolingües de frecuencias se publicaron en las décadas de 1980 y 1990, a saber, el *Xiandai Hanyu Changyongci Cipin Cidian: Yinxu Bufen* (现代汉语常用词词频词典, 音序部分) ‘Diccionario de frecuencias de palabras frecuentes del chino por orden alfabético’ (Liu et al, 1990), el *Hanyu Cihui de Tongji yu Fenxi* (汉语词汇的统计与分析) ‘Estadísticas y análisis del léxico chino’ (Beijing Yuyan Xueyuan Yuyan Jiaoxue Yanjiusuo, 1985) y el *Xiandai Hanyu Pinlü Cidian* (现代汉语频率词典) ‘Diccionario de frecuencias del chino moderno’ (Beijing Yuyan Xueyuan Yuyan Jiaoxue Yanjiusuo, 1986). *A Frequency Dictionary of Mandarin Chinese*, de Xiao et al. (2009) es el más reciente e incluye las 5.000 palabras y 2.000 caracteres chinos de uso más frecuente. Todas las entradas y ejemplos de uso se presentan con sus equivalentes al inglés. El diccionario se basa en varios corpus, como el *Lancaster Corpus of Mandarin Chinese (LCMC)* (10), el *UCLA Corpus of Written Chinese* (29) y el *Lancaster Los Angeles Spoken Chinese Corpus (LLSCC)* (11), entre otros, y constituye una valiosa herramienta para los estudiantes de CLE cuya lengua materna es el inglés. Sin embargo, el *Xiandai Hanyu Cidian* (现代汉语词典) ‘Diccionario del chino moderno’ (Instituto de Lingüística de la Academia China de Ciencias Sociales, Oficina de Edición del Diccionario, 2012), que tiene el estatus de diccionario normativo en China, aún tiene pendiente la creación de un corpus de referencia para su actualización metodológica. Liu y Zhang (2015) analizaron su sexta edición con la ayuda de varios corpus chinos importantes, incluyendo el *CCL Corpus of Chinese Texts* (25) de la Peking University, el *7 the National Language Committee of China (CNC)* (16) y el *Media Language Corpus* (7) de la Communication University of China, con el fin de identificar las áreas temáticas que necesitaban ser mejoradas. Sus resultados apuntaron a que los corpus chinos existentes no pueden satisfacer las necesidades de compilación de este diccionario de referencia debido a sus limitaciones, obsolescencia y estilo discursivo, por lo que señalaron la necesidad urgente de elaborar un gran corpus multimodal y abierto para llevar a cabo dicha empresa, tarea que sigue aún pendiente.

3.2.2. Diseños curriculares y elaboración de materiales docentes

El uso de corpus en el diseño curricular y elaboración de materiales docentes se concreta principalmente en su creación y revisión.

Por un lado, los corpus nos proporcionan una gran variedad de ejemplos lingüísticos de uso real que nos permiten observar aspectos sintácticos, coocurrencias, así como las diferencias sutiles entre palabras sinónimas o semánticamente cercanas. En esta línea, Qi (2016) explica la diferencia entre los términos *xinghao* (幸好), *kuide* (亏得) y *haozai* (好在), todos traducibles por ‘afortunadamente’, basándose en ejemplos del *CCL corpus*, y sugiere que su enseñanza puede partir de aspectos semánticos, contextuales, de carga afectiva y del tipo de oración condicional en el que aparecen. Otro ejemplo reciente lo encontramos en *Claves de la gramática china* (Fisac, 2023), donde la autora explica que se ha servido de los corpus *CNC* (16) y *CCL* (25) como fuentes para encontrar ejemplos de uso de los puntos gramaticales que aborda en su obra.

Por otro lado, los corpus pueden contribuir al diseño curricular y nos pueden ayudar a renovar las metodologías y los materiales vigentes. Por ejemplo, la información sobre frecuencias que proporcionan puede usarse para ajustar el contenido de nuestra docencia ya que, tal y como comenta Mindt (1996: 245-246), cualquier plan de estudios debe elaborarse sobre una base empírica y no intuitiva, por lo que la frecuencia de uso puede servir de guía para determinar las prioridades docentes. En este sentido, contamos con la ‘Lista de caracteres estandarizados frecuentes’ (通用规范汉字表, *Tongyong guifan hanzi biao*), que se elaboró en base a la frecuencia de uso de los caracteres chinos en nueve corpus distintos, entre los que

destaca el CNC (16). Aunque se trata de un corpus equilibrado general de acceso limitado, su información sobre frecuencias se encuentra en abierto.⁶ Según la información proporcionada, podemos observar que los primeros 1.000 caracteres cubren el 89,6% de los caracteres que aparecen en un texto general, mientras que los primeros 2.500 cubren ya un 98,53%, por lo que estos caracteres se consideran el núcleo de la enseñanza del chino, ya sea como lengua materna o extranjera.

Dada su relevancia, esta lista constituye una referencia importante tanto para la selección de caracteres y de vocabulario en la enseñanza del CLE, como para evaluar la idoneidad de determinados materiales didácticos, aunque también hay que tener en cuenta la necesidad de su adaptación al perfil de los estudiantes. En resumen, la información sobre la frecuencia de los caracteres y las palabras que nos proporcionan los corpus constituye una referencia valiosa no solamente para mejorar los materiales didácticos existentes, sino también para elaborar nuevos recursos docentes.

3.2.3. *Aprendizaje guiado por datos*

El uso de corpus aporta ideas interesantes sobre cómo enseñar y ofrece enfoques y soluciones pedagógicas diferentes de la enseñanza tradicional. Johns (1991: 2) subraya que las concordancias de los corpus tienen importantes implicaciones para la enseñanza de idiomas, argumentando que han llevado la enseñanza a la era del aprendizaje guiado por datos (DDL), basado en la observación, la clasificación y la generalización. Por su parte, Carter y McCarthy (1995: 155) afirman que el uso de corpus en la enseñanza de la gramática de la lengua hablada ha conducido a una metodología docente más exploratoria, convirtiéndose en lo que llaman las 3Is (ilustración-interacción-inducción).

Por su parte, Cowie (2006: 579) afirma que “native-like proficiency of a language depends crucially on knowledge of a stock of prefabricated units”. Es decir, en su opinión, el uso de bloques prefabricados de palabras (*chunks*) acelera el proceso de comprensión y producción lingüística.

Desde nuestro punto de vista, la información cuantitativa de las colocaciones y concordancias, presentadas habitualmente en forma de *key words in context* (KWIC) ofrece la posibilidad de identificar dichos bloques prefabricados en chino, lo que puede ayudarnos a prestar atención a las estructuras previas y posteriores de una determinada palabra, las estructuras con alta frecuencia de aparición y a generalizar sobre el uso de las estructuras relacionadas en la enseñanza del vocabulario y la gramática.

Un ejemplo ilustrativo es Tian (2006), quien recurrió a corpus para cuantificar y analizar las palabras utilizadas antes y después de la expresión *you dian* (有点) ‘un poco’, lo cual le permitió resumir sus características semánticas e identificar las categorías gramaticales a las que corresponden. Su objetivo era que los resultados de este estudio basado en corpus ayudasen a los profesores a enseñar un chino auténtico y a evitar la creación de estructuras siguiendo solo reglas gramaticales.

De manera similar, Liu y Hu (2023) también afirman que utilizar un corpus propio para la enseñanza de léxico chino facilita la discriminación de sinónimos y la colocación de palabras para un aprendizaje más exploratorio y autónomo con mejores resultados.

A partir de los resultados de estos dos estudios, nos gustaría destacar la utilidad del *Academia Sinica Balanced Corpus of Modern Chinese* (9), que es un corpus abierto de procesamiento profundo que puede ser un buen recurso para la creación de materiales. Para realizar búsquedas más precisas, se pueden seleccionar palabras clave o palabras reduplicadas y categorías gramaticales simultáneamente, tal como muestra la Figura 2.

⁶ Véase: <https://github.com/bedlate/cn-corpus>.

中央研究院平衡語料庫

請輸入你所要查詢的關鍵字

關鍵詞或重疊詞	詞類	特徵
想 <input type="text"/> 或 <input type="text"/>	VC (動作及物動詞) <input type="text"/>	p1 (合併中插-前) <input type="text"/>

Figura 2: Ejemplo de búsqueda básica en el Academia Sinica Balanced Corpus of Mordern Chinese

Además, es posible activar más filtros en la búsqueda avanzada, de modo que se pueden satisfacer una gran variedad de necesidades (ver la Figura 3).

☐ 縮小查詢範圍

■ 設定文類搜尋範圍:

☐ 報導 ☐ 散文 ☐ 評論 ☐ 詩歌 ☐ 信函 ☐ 會話 ☐ 演講 ☐ 語錄 ☐ 劇本
☐ 公告啟事 ☐ 說明手冊 ☐ 傳記日記 ☐ 會議記錄 ☐ 小說故事寓言 ☐ 廣告或圖文

■ 縮小主題範圍:

☐ 戲 ☐ 思想 ☐ 心理 ☐ 宗教 ☐ 數學 ☐ 天文 ☐ 物理 ☐ 化學 ☐ 礦冶 ☐ 生物 ☐ 地理 ☐ 環保 ☐
☐ 工程 ☐ 經濟 ☐ 財政 ☐ 商管 ☐ 行銷 ☐ 政黨 ☐ 內政 ☐ 軍事 ☐ 司法 ☐ 教育 ☐ 文化 ☐ 歷史 ☐ 民族
☐ 語文 ☐ 傳播 ☐ 公益 ☐ 福利 ☐ 人事 ☐ 犯罪 ☐ 災禍 ☐ 音樂 ☐ 舞蹈 ☐ 雕塑 ☐ 美術 ☐ 攝影 ☐
☐ 技藝 ☐ 文物 ☐ 建築 ☐ 旅遊 ☐ 體育 ☐ 食物 ☐ 醫療 ☐ 衣飾 ☐ 影藝 ☐ 人物 ☐ 訊息 ☐ 消費 ☐ 家庭
☐ 政治學 ☐ 衛生保健 ☐ 兒童文學 ☐ 俠義文學 ☐ 言情文學 ☐ 交通運輸 ☐ 文學理論 ☐ 鄉土文學 ☐ 農
☐ 漁牧業 ☐ 政治現象 ☐ 國家政策 ☐ 國際關係 ☐ 藝術總論 ☐ 社會現象 ☐ 大學科學 ☐ 統計調查 ☐ 批評與
☐ 鑑賞 ☐ 其他文學創作

☐ 排序

依 的 排序

依 的 排序

依 的 排序

☒ 詞類統計 查詢範圍 -5 到 5 > 0: 關鍵詞右邊 < 0: 關鍵詞左邊 距離不能超過10
☐ Collocation 查詢範圍 -5 到 5 > 0: 關鍵詞右邊 < 0: 關鍵詞左邊 距離不能超過10
 對象為 詞及詞類 Mi值 排序: 互見訊息 出現詞頻 =>

☐ 過濾 過濾範圍 -10 到 10 > 0: 關鍵詞右邊 < 0: 關鍵詞左邊 -10<= 範圍 <=10
☐ 包含
 條件一: 包含關鍵詞 或 詞類 特徵
 或包含關鍵詞 或 詞類 特徵

Figura 3: Búsqueda avanzada en el Academia Sinica Balanced Corpus of Mordern Chinese

En definitiva, acceder a una gran cantidad de material lingüístico auténtico y ordenado nos permite observar el contexto en el que aparece una palabra y comprender mejor su significado, lo que facilita su enseñanza y puede redundar en la naturalidad de la producción lingüística. Tal como sugiere Johns (1991: 4-5), los corpus pueden adaptarse a distintos niveles de aprendizaje y pueden ser usados tanto por el profesorado como por el alumnado. De este modo, nuestros estudiantes pueden cambiar de identidad, empoderarse y tomar las riendas de su aprendizaje, asumiendo el papel de investigadores y accediendo a datos lingüísticos directamente consultando un corpus adecuado.

3.2.4. Evaluación de la competencia lingüística

Alderson (1996: 258-259) señaló las ventajas de utilizar datos de corpus como base para juzgar las capacidades, los conocimientos y el rendimiento de los examinandos. También Ball (2001: 7) y Hunston (2002: 205) señalaron que las principales organizaciones de evaluación han estado utilizando corpus para diferentes fines: 1) archivar plantillas de examen; 2) desarrollar materiales de evaluación; 3) optimizar los procedimientos de evaluación; 4) mejorar la calidad de la evaluación; 5) validar exámenes; y 6) estandarizar pruebas.

Por otro lado, no podemos subestimar el uso de corpus de estudiantes ya que, por ejemplo, pueden ayudarnos a determinar la influencia de las estructuras en la dificultad de las pruebas (Ball, 2001: 7), a validar su fiabilidad y validez utilizando datos reales (He y Dai, 2006: 393) o, simplemente, ofrecernos datos objetivos para llevar a cabo investigaciones sobre pruebas de idiomas (Biber et al., 2004: v).

Gracias a nuestra revisión bibliográfica, hemos constatado que el uso actual de los corpus en la enseñanza del CLE es muy escaso, por lo que las experiencias exitosas del TESL (Teaching English as a Second Language) nos pueden servir de ejemplo. En este sentido, en el caso concreto de los profesores de CLE aconsejamos utilizar los corpus *CCL* (25), *CNC* (16), *Academia Sinica Balanced corpus* (9) y el corpus *BCC* (5) de la Beijing Language and Culture University, que es el mayor de todos, porque son de nativos, equilibrados y de acceso abierto. Estos corpus se pueden utilizar como herramientas auxiliares para elaborar listas de vocabulario para distintos niveles y desarrollar sistemas automáticos de calificación, ya que su elaboración no se basa simplemente en la experiencia de los lingüistas, sino también en el uso real de la lengua (Hazenbergh y Hulstun, 1996: 146). Basándose en el corpus de Composición Dinámica HSK (3), Rao y He (2024) propusieron un modelo de puntuación automática que integra características de errores interlingüísticos con rasgos semánticos a nivel de palabra de un modelo lingüístico entrenado previamente. Los resultados experimentales demostraron que el modelo mejoraba significativamente la consistencia y la estabilidad de las puntuaciones y reflejaba de forma más precisa el nivel real de escritura de los estudiantes, lo que proporciona un respaldo empírico sólido para mejorar la capacidad explicativa y la fiabilidad de los sistemas automáticos de corrección.

Además, hay otras aplicaciones más sencillas y prácticas que no necesitan modelos extra. Por ejemplo, de acuerdo con Wu et al. (2024), la National Academy for Educational Research (NAER) de Taiwán integró los resultados del estándar de niveles de competencia en lengua china y los recursos de los corpus *COCT* (14) para desarrollar diversas herramientas didácticas, entre las que se encuentra el *Zuowen cuobiezi zidong pigai xitong* (作文错别字自动批改系统) ‘Sistema de corrección automática de errores ortográficos en redacción’, cuyo objetivo es mejorar la automatización y la eficacia de la retroalimentación en la enseñanza del chino, así como potenciar la función de evaluación diagnóstica dentro del proceso educativo.

3.2.5. Formación del profesorado

El papel que desempeña el profesorado en el aprendizaje de idiomas de sus alumnos hace que su capacidad para usar los corpus afecte también indirectamente la metodología y la calidad de su práctica docente, permitiendo que los alumnos se beneficien de las ventajas que ofrecen. De acuerdo con O’Keeffe y Farr (2003: 412), es necesario incluir la lingüística de corpus en la formación del profesorado de idiomas para mejorar su conciencia lingüística y su competencia digital e investigadora. Un profesor que domina la lingüística de corpus es un docente empoderado que puede crear sus propios materiales y disponer de una herramienta de consulta más sobre la producción lingüística real y actualizada de los hablantes nativos, recurso que puede ser un buen complemento a los manuales, que suelen quedar obsoletos en este sentido.

La formación en línea y presencial sobre corpus para profesores de lengua china alrededor del mundo es limitada, no obstante, podemos encontrar recursos de apoyo. Por ejemplo, Liu (2020-2023) publicó nueve volúmenes en la colección *Yuyan Jisuan yu Zhineng Hanyu Jiaoxue* (语言计算与智能汉语教学) ‘Computación lingüística y enseñanza inteligente de la lengua china’ con la editorial *Foreign Language Teaching and Research Press*,⁷ donde ofrece reflexiones teóricas, recursos y metodologías docentes sobre la práctica y la investigación de la enseñanza de la lengua china mediante corpus y métodos lingüísticos computacionales. Por su parte, Tao y Chen (2019) recopilaron 17 artículos sobre las cuestiones pedagógicas relacionadas con la enseñanza del chino para fines específicos o profesionales con la esperanza de promover la docencia y la investigación en este ámbito. Este conjunto de contribuciones está dirigido a quienes se dedican a la enseñanza del CLE y cubre sus necesidades de formación en corpus, por lo que representa una contribución interesante en el fomento de la aplicación de los corpus en la enseñanza y la investigación de la lengua china.

3.3. Utilidad de los corpus orientados a la enseñanza

El uso pedagógico de los corpus se materializa en la elaboración y aplicación de corpus destinados a la enseñanza. De acuerdo con Leech (1997) pueden ser de lengua para fines específicos, de adquisición de la L1 y la L2, bilingües o multilingües. A continuación, explicaremos con más detalle en cada uno de ellos, a excepción de los corpus de adquisición de la L1, ya que la adquisición de la lengua materna está fuera de nuestro objeto de estudio.

3.3.1. Corpus de lengua para fines específicos

La lengua para fines específicos (LFE) ocupa un lugar importante entre los objetivos de la enseñanza de idiomas (Leech, 1997: 18). Ya sea por motivos de estudio o de trabajo, a menudo es necesario conocer las especificidades y características de una lengua de un campo de especialidad. En niveles avanzados de enseñanza de la lengua china también se suelen introducir asignaturas de chino para fines específicos, momento en el que se harán patentes las ventajas y la necesidad de recurrir a los corpus de este tipo. Debido a que el número de corpus de LFE multilingües es limitado, a continuación, presentamos los corpus chinos temáticos existentes en dos grandes categorías: los monolingües y los bilingües.

En cuanto a los corpus monolingües, el *New Era People’s Daily Segmented Corpus (NEPD)* (13), elaborado por la Nanjing Agricultural University, está enfocado a los medios de comunicación y cuenta con una gran difusión. Por otra parte, la serie de corpus creados por la National Taiwan Normal University (20-24), accesibles a través de su plataforma, abarcan muchos ámbitos temáticos de interés, incluyendo mundo académico, derecho, economía, ciencia, turismo y subtitulado, y también se pueden usar para la investigación y la enseñanza de la lengua oral.

En el caso de los corpus bilingües, los más conocidos son el *Zhongguo Falü Fagui Han-Ying Pingxing Yuliaoku* (中国法律法规汉英平行语料库) ‘Corpus paralelo de documentos legales de China’ (27) de la Shaoxing University, el *Chinese-English / English-Chinese translation Corpus of signs* (2) de la Beijing International Studies University y el *English-Chinese Medical Parallel Corpus* (26) de la Shanghai Jiaotong University, dedicados a temas jurídicos, carteles públicos y a la medicina, respectivamente.

Estos corpus temáticos, a la vez que proporcionan recursos terminológicos para la enseñanza de la lengua china para fines específicos, también pueden facilitar nuestro análisis de las características discursivas en ámbitos de especialidad del chino, como, por ejemplo, las palabras más frecuentes y sus peculiaridades de uso en esta área, las características de

⁷ Véase: <http://www.language-tech.cn/tcsl/>.

las categorías léxicas utilizadas o la longitud media de las oraciones. Asimismo, los corpus temáticos bilingües también ofrecen recursos didácticos para la enseñanza de la traducción y pueden ofrecer propuestas para la traducción o interpretación especializada entre pares de lenguas. Tal como indica Sánchez-Gijón (2004: 136-142), todas las unidades de conocimiento especializado, ya sea terminología, colocaciones, usos verbales o locuciones, entre otros, son detectables para los traductores a través de corpus y pueden contribuir a la comprensión y, por ende, a la traducción apropiada en la lengua meta.

3.3.2. *Corpus de adquisición del chino como lengua extranjera*

De acuerdo con Selinker (1972), la interlengua es una lengua intermedia entre la lengua materna y la lengua meta producida por los alumnos de una lengua extranjera (LE) o de una segunda lengua (L2). Por lo tanto, los corpus de adquisición de L2 se consideran corpus monolingües de interlengua y se utilizan principalmente para la enseñanza de la LE. Pueden clasificarse en cuatro tipos: orales, escritos, mixtos y multimodales. Su anotación es diferente a la de otros tipos de corpus ya que, aparte de las propiedades léxicas y los significados, se etiquetan también los distintos errores de uso, sean estos ortográficos, gramaticales o pragmáticos.

Los corpus chinos de interlengua ampliamente utilizados como el *Chinese Interlanguage Text Corpus* (4) y el *COCT Zhongjie Yuliaoku* (COCT中介语料库) ‘COCT Corpus de interlengua’ (15) se centran en los errores de redacción por parte de estudiantes de CLE. El *HSK Dynamic Composition Corpus* (3) y el *TOCFL Learner Corpus* (17) están centrados en los errores de redacción en los exámenes oficiales de chino. En cambio, el *Guangwai-Lancaster Chinese Learner Corpus* (8), el *Chinese Learners Written Corpus* (19) y el *Chinese Learners Spoken Corpus* (18) reúnen errores tanto escritos como orales. Por último, el *Chinese Interlanguage Text Corpus* (4) es un corpus multimodal que contiene una amplia gama de errores detectados en las actividades de práctica lingüística e incluye sonidos, imágenes y lenguaje no verbal, entre otros. Los corpus de estudiantes se utilizan actualmente para el diseño curricular, la elaboración de materiales docentes y la investigación en metodología docente. Tal y como señala Meunier (2002: 123), para orientar el diseño de los planes de estudio y materiales para la enseñanza, no basta limitarse a la información de frecuencia que proporcionan los corpus de hablantes nativos ya que se debe alcanzar un equilibrio entre la información sobre la frecuencia, el nivel de dificultad y la pertinencia pedagógica.

En este tipo de corpus podemos detectar los errores más frecuentes que cometen los estudiantes y las diferencias de uso entre los estudiantes y los hablantes nativos, contribuyendo así a predecir los posibles errores provocados por la transferencia negativa y a optimizar la docencia. Por ejemplo, comparando el *HSK Dynamic Composition Corpus* (3) con el *BFSU ToRCH Chinese Corpora* (1), Li (2023) identifica los tipos de estructura y aspectos semánticos en los que se debe hacer hincapié durante la enseñanza de las oraciones con *ba* (把). Guo (2023) realizó un estudio en profundidad basado en el uso de la palabra *dui* (对) a partir de tres tipos de corpus: corpus de chinos nativos, corpus de libros de texto de CLE y corpus de estudiantes de chino. Evaluó la representatividad del idioma de los libros de texto, las características de adquisición de los alumnos y los tipos de sesgos, y sugirió que los profesores podrían optimizar sus contenidos y estrategias de enseñanza con la ayuda de los datos del corpus para lograr un diseño didáctico más específico.

Además, dado que estos corpus suelen especificar el nivel de los estudiantes, nos pueden ayudar a comprender las características de su aprendizaje y de su interlengua en las diferentes etapas, lo que nos puede ayudar a anticiparnos a los errores y orientarnos a la hora de preparar nuestras clases adaptándolas mejor a las dificultades y necesidades particulares de cada perfil de alumno.

3.3.3. *Corpus bilingües o multilingües*

Los corpus bilingües o multilingües a los que se hace referencia en este apartado son específicamente aquellos en los que los textos en varias lenguas son comparables entre sí hasta cierto punto, es decir, los que suelen denominarse *paralelos* (平行, *pingxing*). Puesto que la finalidad es utilizar los corpus para la enseñanza del CLE y no para analizar las características del chino traducido, los datos relacionados con este último no deberían de tenerse en cuenta.

La alineación de los textos permite la visualización de unidades textuales equivalentes en diferentes lenguas (Leech, 1997: 21-22). Los corpus paralelos comparables no solo constituyen un recurso clave para el desarrollo de la lingüística contrastiva, la pragmática y la traducción automática, sino que también pueden usarse en la enseñanza de la traducción y en su investigación. En este sentido, destaca el trabajo de Ma (2022) que, con la ayuda del *Parallel Corpus Chinese-English of A Dream of Red Mansions* (28) de la Shaoxing University, analizó la traducción de tres adverbios que indican un cambio en las expectativas del hablante, a saber, *fan* (反), *jing* (竟) y *pian* (偏), para identificar tanto los aspectos semánticos comunes como sus particularidades de uso. En definitiva, creemos que este tipo de estudios pueden hacer una aportación interesante tanto a la enseñanza e investigación en traducción, como a la enseñanza del CLE.

4. CONCLUSIONES

A pesar de que el desarrollo de los corpus de chino, especialmente orientados a la enseñanza, ha brillado por su ausencia en los países hispanohablantes, en los últimos años han surgido algunas iniciativas interesantes que vale la pena mencionar. Por una parte, la Universidad de Valladolid coordina el proyecto *PaCorES*, que incluye el *Parallel Corpus Chinese<>Spanish, PaCheS* (31), diseñado como herramienta para la traducción y la enseñanza de lenguas. Por otra parte, existen iniciativas centradas en la lengua oral, como el *COLHACH* (12), impulsado por la Universidad Complutense de Madrid y, finalmente, el *Corpus C-ORAL CHINA* (30), desarrollado por la Universidad Autónoma de Madrid, ambos dedicados al estudio del chino hablado a partir de datos reales. Aunque aún en fases iniciales, estos corpus representan un avance significativo en este campo. Sin embargo, la aplicación didáctica de corpus en la enseñanza del CLE para hispanohablantes y su investigación se encuentra todavía en una etapa poco desarrollada. Los corpus existentes proceden principalmente de China y sus principales usuarios son profesores nativos. Eso se debe a que los profesores de CLE de los países hispanohablantes suelen ignorar la existencia de estos corpus, desconocen su potencial para la docencia o carecen de la formación y competencias necesarias para explotarlos para su docencia.

Puesto que existen diferencias significativas entre el chino y el español en cuanto al orden oracional y la estructura discursiva, dos rasgos lingüísticos clave que puede revelar el análisis de corpus, creemos que potenciar la investigación sobre el uso de corpus en la enseñanza del CLE y promover la traslación de sus resultados a la práctica docente puede contribuir a mejorar la calidad de la enseñanza. Por ello, en este trabajo nos hemos propuesto, en primer lugar, hacer una selección y análisis de una muestra representativa de corpus de chino, con el fin de ofrecer una visión panorámica de los corpus existentes y de sus principales características y, en segundo lugar, revisar las posibilidades didácticas del uso de corpus en la enseñanza del CLE a partir de estudios previos y experiencias ya documentadas para ilustrar los enfoques teóricos con casos de aplicación reales con la esperanza de animar al profesorado a llevar dichas aplicaciones a su propia práctica en el aula.

A partir de la metodología establecida, se ha creado una base de datos consultable en acceso abierto con 56 corpus de chino, categorizados en función de diferentes parámetros de análisis. La creación de esta base de datos cumple con el primer objetivo de este estudio y ofrece a la comunidad académica una lista organizada de recursos que se encuentran dispersos y a menudo son de difícil acceso, lo cual constituye una aportación en sí misma. Con el fin de apuntar las posibles aplicaciones de estos corpus en la enseñanza del CLE, se han presentado de forma sistemática ejemplos concretos de estas aplicaciones de acuerdo con la clasificación de Leech (1997). De este modo, se ha partido de los usos directos, reservados para estudiantes de niveles avanzados, y se han revisado experiencias existentes de los usos indirectos, como la confección de diccionarios, la elaboración de listas de palabras a partir de datos estadísticos de frecuencias, la creación de materiales docentes y de herramientas de evaluación, así como la elaboración de corpus orientados a la enseñanza de la lengua o la traducción. Estos últimos pueden servir para caracterizar lenguajes de especialidad, describir determinados puntos gramaticales, buscar equivalencias entre lenguas e incluso identificar errores procedentes de la interlengua de los estudiantes de diferentes perfiles y niveles de aprendizaje. En el contexto de la didáctica del CLE se han recomendado particularmente dos corpus para la enseñanza del CLE, a saber, el Academia Sinica Balanced Corpus of Modern Chinese (CCL) y el corpus de la Beijing Language and Culture University (CNC).

Toda investigación presenta limitaciones y este trabajo no es una excepción. Así, por ejemplo, no se ha podido presentar un análisis en profundidad de los 56 corpus de la muestra por motivos de espacio y porque este no eran el objetivo del presente estudio. Tampoco se han explorado aplicaciones propias de los corpus, sino que se ha optado por documentar las de otros investigadores. Esto se debe a que una experiencia docente de investigación-acción con el uso de corpus podría constituir un estudio en sí mismo, lo cual tampoco constaba entre los objetivos de este artículo. En definitiva, en este sentido, pensamos que se han abierto numerosas líneas de investigación, tanto para nosotras como para el resto de los investigadores, ya sea realizando un análisis contrastivo de los diferentes corpus de la base de datos recopilada, ya sea realizando estudios empíricos centrados en aplicaciones concretas de los corpus en la enseñanza del CLE o desarrollando corpus para o de aprendices hispanohablantes de CLE.

En definitiva, el carácter exploratorio del presente trabajo pretendía ser una fuente de inspiración para aquellos que quieran explorar el uso de los corpus en el aula del CLE y no sepan por dónde empezar ante la cantidad y diversidad de la oferta. Conocer las características y posibles aplicaciones de los recursos existentes puede ser un buen punto de partida y facilitar que se avance en esta línea. Como docentes debemos esforzarnos en desarrollar y actualizar constantemente nuestra literacidad digital que, en este caso, pasa por conocer los corpus disponibles, sus puntos fuertes y débiles, así como saber cómo interrogarlos para obtener información relevante para la docencia y la investigación. Este trabajo no cubre todas estas dimensiones, pero sí hace una pequeña contribución en presentar de forma ordenada y sistemática las opciones a nuestro alcance a partir de las cuales, el profesorado interesado puede empezar a investigar o experimentar por su cuenta en función de sus propias necesidades.

AGRADECIMIENTOS

Los resultados de esta investigación se han beneficiado del apoyo del Ministerio de Ciencia, Innovación y Universidades de España a través del proyecto PID2024-156763OB-I00 (GENE-RATESE-2LT), y han contado con el respaldo del grupo de investigación GELEA2LT.

Las autoras del presente trabajo quieren agradecer a sus compañeros del grupo de investigación todos sus comentarios a una versión preliminar de este trabajo. Asimismo, también queremos agradecer a los revisores anónimos de la revista su atenta lectura y valiosas sugerencias que, sin duda, han contribuido a mejorar la calidad final del artículo.

REFERENCIAS

- Alderson, C. (1996). Do corpora have a role in language assessment? En J. Thomas & M. Short (Eds.), *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech* (pp. 248-259). Longman.
- Ball, F. (2001). Using corpora in language testing. *Cambridge: Cambridge ESOL*, 6, 6-8.
- Beijing Yuyan Xueyuan Yuyan Jiaoxue Yanjiusuo (北京语言学院语言教学研究所). (1985). *Hanyu Cihui de Tongji yu Fenxi* (汉语词汇的统计与分析) [Análisis estadístico del léxico del chino]. Waiyu Jiaoxue yu Yanjiu Chubanshe.
- Beijing Yuyan Xueyuan Yuyan Jiaoxue Yanjiusuo (北京语言学院语言教学研究所). (1986). *Xiandai Hanyu Pinlü Cidian* (现代汉语频率词典) [Diccionario de frecuencias del chino moderno]. Beijing Yuyan Xueyuan Chubanshe.
- Belz, J., & Vyatkina, N. (2005). Learner corpus analysis and the development of L2 pragmatic competence in networked inter-cultural language study: The case of German modal particles. *Canadian Modern Language Review*, 62, 17-48. <https://doi.org/10.3138/cmlr.62.1.17>
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511804489>
- Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing Language Use in the University: Analysis of the TOEFL 2000 spoken and Written Academic Language Corpus*. Educational Testing Service.
- Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning*, 67(2), 348-393. <https://doi.org/10.1111/lang.12224>
- Carter, R., & McCarthy, M. (1995). Grammar and the Spoken Language. *Applied Linguistics*, 16(2), 141-158. <https://doi.org/10.1093/applin/16.2.141>
- Conrad, S. (2000). Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century? *TESOL Quarterly*, 34(3), 548-560. <https://doi.org/10.2307/3587743>
- Cowie, A. (2006). Phraseology. En K. Brown (Ed.), *Encyclopedia of Language & Linguistics* (2.^a ed., Vol. 6, pp. 579-585). Pergamon Press. <https://doi.org/10.1016/B0-08-044854-2/00440-5>
- Farr, F. (2008). Evaluating the Use of Corpus-based Instruction in a Language Teacher Education Context: Perspectives from the Users. *Language Awareness*, 17(1), 25-43. <https://doi.org/10.2167/la414.0>
- Fisac, T. (2023). *Claves de la gramática china* (Vol. 35). Bellaterra Edicions.
- Gong, Y. (2019). Las nominalizaciones como estrategia discursiva en el discurso político: Análisis comparado entre el español y el chino. *Círculo de Lingüística Aplicada a la Comunicación*, 79, 257-276. <https://doi.org/10.5209/clac.65658>
- Guo, J. (2023). *Usage-based Analysis of Schematic and Multifunctional Verb Argument Constructions in Three Contrasting Corpora: A Native Chinese Corpus, a CFL Textbook Corpus and a Chinese Learner Corpus*. <https://www.repository.cam.ac.uk/handle/1810/362234>

Hazenbergh, S., & Hulstun, J. H. (1996). Defining a Minimal Receptive Second-Language Vocabulary for Non-native University Students: An Empirical Investigation. *Applied Linguistics*, 17(2), 145-163. <https://doi.org/10.1093/applin/17.2.145>

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing*, 23(3), 370-401. <https://doi.org/10.1191/0265532206lt333oa>

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139524773>

Instituto de Lingüística de la Academia China de Ciencias Sociales, Oficina de Edición del Diccionario. (2012). *Xiandai Hanyu Cidian* (现代汉语词典) [Diccionario del chino moderno] (6.^a ed). Commercial Press.

Johns, T. (1991). Should you be persuaded. Two samples of data-driven learning materials. *English Language Research Journal*, 4, 1-16.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Routledge.

Leech, G. (1997). Teaching and Language Corpora: A Convergence. En A. Wichmann, S. Fligelstone, McEnery, T., & Knowles, Gerry (Eds.), *Teaching and Language Corpora* (pp. 1-23). Longman. <https://doi.org/10.4324/9781315842677-1>

Li, X. (李雪婷). (2023). Duiwai Hanyu Jiaoxue zhong “Ba”ziju Jiaoxue de Zhongdian yu Nandian——Jiyu Ziranyu Yuliaoku yu Zhongjieyu Yuliaoku de Duibi Yanjiu (对外汉语教学中“把”字句教学的重点与难点——基于自然语料库与中介语语料库的对比研究) [Puntos clave y difíciles de la enseñanza de las oraciones “ba” en chino como lengua extranjera: Un estudio comparativo basado en el corpus de lengua natural y el de interlingua]. *Beijing Huagong Daxue Xuebao: Shehui Kexue Ban* (北京化工大学学报: 社会科学版) / *Journal of Beijing University of Chemical Technology*, 4(125), 108-115.

Liang, M. (梁茂成), Li, W. (李文中), & Xu, J. (许家金). (2010). *Yuliaoku Yingyong Jiaocheng* (语料库应用教程) [Manual Práctico de Corpus]. Waiyu Jiaoxue yu Yanjiu Chubanshe.

Liu, H. (刘华). (2020-2023). *Yuyan Jisuan yu Zhineng Hanyu Jiaoxue* (语言计算与智能汉语教学) [Lingüística computacional y enseñanza inteligente del chino] (9 vol.). Waiyu Jiaoxue yu Yanjiu Chubanshe.

Liu, Y. (刘永静), y Hu, X. (胡宪丽). (2023). Jiyu Rongmeiti Yuliaoku de Guoji Zhongwen Cihui Jiaoxue Anli Yanjiu (基于融媒体语料库的国际中文词汇教学案例研究) [Estudio de caso sobre la enseñanza internacional de vocabulario chino basado en el corpus de convergencia de medios de comunicación]. *Hanyu Guoji Jiaoyu Xuebao* (汉语国际教育学报) / *Journal of International Chinese Education*, 14, 153-166.

Liu, Z. (刘泽权), y Zhang, D. (张丹丹). (2015). *Xiandai Hanyu Cidian Di-liu ban zhi Yizhu jian Lun Jiyu Yuliaoku de Hanyu Cidian Bianzuan de Biyaoxing* (《现代汉语词典》第六版之遗珠兼论基于语料库的汉语词典编纂的必要性) [Some Problems of *The Contemporary Chinese Dictionary* (6th Ed.): with a Call for a Corpus-Based Chinese Lexicography]. *Journal of Chinese Linguistics*, 43(1), 196-214.

Liu, Y. (刘源), Liang, N. (梁南元), Wang, D. (王德进), Zhang, S. (张社英), Yang, T. (杨铁鹰), Jie, C. (揭春雨), y Sun, W. (孙伟). (1990). *Xiandai Hanyu Changyongci Cipin Cidian: Yinxu Bufen* (现代汉语常用词词频词典, 音序部分) [Diccionario de frecuencias de palabras de uso frecuente del chino moderno por orden alfabético]. Yuhang Chubanshe.

Ma, W. (马威艳). (2022). Jiyu ‘Hongloumeng’ Shuangyu Pingxing Yuliaoku de Hanyu ‘Fan’ ‘Jing’ ‘Pian’ Fan Yuqi Biaoji Shenceng Yuyi Fenxi (基于《红楼梦》双语平行语料库的汉语“反”“竟”“偏”反预期标记深层语义分析) [Análisis semántico de los marcadores chinos anti-expectativos “fan (反)”, “jing (竟)” y “pian (偏)” basado en un corpus paralelo bilingüe del Sueño en el pabellón rojo]. *Liaoning Gongye Daxue Xuebao. Shehui Kexue Ban* (辽宁工业大学学报社会科学版), 24(5), 67-70.

McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. En Hinkel, Eli (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 364-380). Routledge. https://www.academia.edu/2654860/What_corpora_can_offer_in_language_teaching_and_learning

McEnery, T., Xiao, R., & Tono, Y. (2010). *Corpus-based language studies: An advanced resource book* (Reprinted). Routledge.

Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. En S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 119-141). John Benjamins Publishing Company. <https://doi.org/10.1075/llt.6.10meu>

Mindt, D. (1996). English corpus linguistics and the foreign language teaching syllabus. En J. Thomas & M. Short (Eds.), *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech* (pp. 232-247). Longman.

Nie, L. (En prensa). El impacto de la globalización y la sinización de los anglicismos en el chino de Hong Kong y Macao en las noticias de entretenimiento. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, 51(2).

Nie, L. y Rovira-Esteva, S. (2025). *Base de datos de corpus chinos: Tipologías y características* [Base de datos]. CORA. Repositori de Dades de Recerca. <https://doi.org/10.34810/data2343>

O’Keeffe, A., & Farr, F. (2003). Using Language Corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications. *TESOL Quarterly*, 37(3), 389-418. <https://doi.org/10.2307/3588397>

Qi, T. (齐婷婷). (2016). ‘Xinghao’, ‘Haozai’, ‘Kuide’ Bianxi ji Duiwai Hanyu Jiaoxue (“幸好”“好在”“亏得”辨析及对外汉语教学) [The Discrimination and Chinese’s Teaching for Foreigner of “Fortunately” “Luckily” “Thanks to”]. *Liupanshui Shifan Xueyuan Xuebao* (六盘水师范学院学报) / *Journal of Liupanshui Normal University*, 28(5), 69-72.

Rao, J., & He, Y. (2024). Research on Automated Scoring Method for HSK Essays with Hybrid Features. En R. Liu, L. Wang, F. Bao, Y. Lu, C. Fang, & M. Dong (Eds.), *2024 International Conference on Asian Language Processing (IALP)* (pp. 439-444). IEEE. <https://doi.org/10.1109/IALP63756.2024.10661138>

Sánchez-Gijón, P. (2004). *L’ús de corpus en la traducció especialitzada: Compilació de corpus ad hoc i extracció de recursos terminològics*. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: Universitat Autònoma de Barcelona.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(3), 209-232. <https://doi.org/10.1515/iral.1972.10.1-4.209>

Tao, H., & Chen, H. H.-J. (Eds.). (2019). *Chinese for Specific and Professional Purposes: Theory, Pedagogical Applications, and Practices*. Springer. <https://doi.org/10.1007/978-981-13-9505-5>

Tian, H. (田宏梅). (2006). Liyong Hanyu Yuliaoku Yanjiu Ciyu Dapei——Yi ‘Youdian’ Wei Li (利用汉语语料库研究词语搭配——以“有点”为例) [Two Major Approaches in Corpus Linguistics to the Study of Collocation: A Case Study with youdian (有点)]. *Jinan Daxue Huawen Xueyuan Xuebao* (暨南大学华文学院学报), 3, 67-73.

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

Pan, S. (2013). *Xin Shidai Han-Ying Da Cidian* (新时代汉英大辞典). *New Age Chinese-English Dictionary* (2nd Edition). Commercial Press. Wu, J. (吴鉴城), Bai, M. (白明弘), y Lin, Q. (林庆隆). (2019). Taiwan Huayuwen Yuliaoku zai Huayuwen Jiaoyu de Yingyong (台湾华语文语料库在华语文教育的应用) / Applying the Corpus of Contemporary Taiwanese Mandarin in Teaching Chinese as a Second Language. *Huayuwen Jiaoxue Yanjiu* (华语文教学研究), 16(3), 29-55. [https://doi.org/10.6393/JCLT.201909_16\(3\).0002](https://doi.org/10.6393/JCLT.201909_16(3).0002)

Wu, X., Yang, Y., Zhou, X., Xia, Y., y Liao, H. (2024). A meta-analysis of interdisciplinary teaching abilities among elementary and secondary school STEM teachers. *International Journal of STEM Education*, 11(38). <https://doi.org/10.1186/s40594-024-00500-8>

Xiao, R., Rayson, P., & McEnery, T. (2009). *A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners*. Routledge. <https://www.routledge.com/A-Frequency-Dictionary-of-Mandarin-Chinese-Core-Vocabulary-for-Learners/Xiao-Rayson-McEnery/p/book/9780415455862>

Zhou, X. (周小兵), Bo, W. (薄巍), Wang, L. (王乐), y Li, Y. (李亚楠). (2017). Guoji Hanyu Jiaocai Yuliaoku de Jianshe yu Yingyong (国际汉语教材语料库的建设与应用) / A Study of the Construction and Application of the TCSOL Material Corpus. *Yuyan Wenzhi Yingyong* (语言文字应用) / *Applied Linguistics*, 101, 125-135.

APÉNDICE 1: CORPUS MENCIONADOS EN ESTE TRABAJO

1. Beijing Foreign Studies University. Xu, J. (许家金), y Wu, L. (吴良平). (2009, 2014, 2019). *BFSU ToRCH family Chinese corpora*. <http://114.251.154.212/cqp>
2. Beijing International Studies University. (2009). *Quanguo Gongshiyu Fanyi Yuliaoku* (全国公示语翻译语料库) [Chinese-English, English-Chinese translation Corpus of Signs].
3. Beijing Language and Culture University. (2006). *HSK Dongtai Zuowen Yuliaoku* (HSK动态作文语料库) [HSK Dynamic Composition Corpus]. <http://hsk.blcu.edu.cn/Login>
4. Beijing Language and Culture University. (2019). *Quanqiu Hanyu Zhongjieyu Yuliaoku* (全球汉语中介语语料库) [Chinese Interlanguage Text Corpus]. <http://qqk.blcu.edu.cn/#/login>

5. Beijing Language and Culture University. Xun, E. (荀恩东). (2016). *BCC Yuliaoku* (BCC 语料库) [Corpus BCC]. <http://bcc.blcu.edu.cn/>
6. Bilingual Dictionary Research Center. Nanjing University. (1997). *NJU_BDRCBC Corpus*. <https://bdrcc.nju.edu.cn/>
7. Communication University of China. (2016). 媒体语言语料库 (MLC). *Media Language Corpus*. <https://ling.cuc.edu.cn/RawPub/>
8. Guangdong University of Foreign Studies, y Lancaster University. (2017). *Guangwai-Lancaster Hanyu Xuexizhe Yuliaoku* (广外—兰卡斯特汉语学习者语料库). *Guangwai-Lancaster Chinese Learner Corpus*. <https://www.sketchengine.eu/guangwai-lancaster-chinese-learner-corpus/>
9. Institute of Information Science, Institute of Linguistics, y Department of Information Technology Services, Academia Sinica. (2010). *Zhongyang Yanjiuyuan Hanyu Biaoji Yuliaoku* (中央研究院现代汉语标记语料库) / *Academia Sinica Balanced Corpus of Modern Chinese*. <http://asbc.iis.sinica.edu.tw/>
10. Lancaster University. Xiao, R. (肖忠华). (2012). *The Lancaster Corpus of Mandarin Chinese version 2 (LCMC)*. <http://114.251.154.212/cqp/lcmc2/>
11. Lancaster University. Xiao, R. (肖忠华), y Tao, H. (陶红印). (2006). *The Lancaster Los Angeles Spoken Chinese Corpus (LLSCC)*. <https://www.lancaster.ac.uk/fass/projects/corpus/LLSCC/index.htm>
12. Li Yue. (2019). *Corpus de Lengua Hablada en Chino (COLHACH)*. Universidad Complutense de Madrid. <https://www.ucm.es/coralingo/corpus-colhach>
13. Nanjing Agricultural University. (2018). *Xin Shidai Renmin Ribao (Fenci) Yuliaoku* (新时代人民日报(分词)语料库) [New Era People's Daily Segmented Corpus (NEPD)]. <https://corpus.njau.edu.cn/>
14. National Academy for Educational Research. (2020). *Taiwan Huayuwen Yuliaoku* (台湾华语语文语料库) [Corpus of Contemporary Taiwanese Mandarin (COCT)]. <https://coct.naer.edu.tw/>
15. National Academy for Educational Research. (2016, 2019). *COCT Zhongjieyu Yuliaoku* (COCT中介语语料库) [Interlingua Learner Corpus (COCT)]. <https://coct.naer.edu.tw/>
16. National Language Committee of China. Xiao, H. (肖航). (1998). *Guojia Yuwei Xiandai Hanyu Yuliaoku* (国家语委现代汉语语料库) [Chinese Corpus by National Language Committee of China (CNC)].
17. National Taiwan Normal University. (2013). *TOCFL Xuexizhe Yuliaoku* (TOCFL学习者语料库) [TOCFL Learner Corpus]. <http://tocfl.itc.ntnu.edu.tw:8080/>
18. National Taiwan Normal University. (2017). *Huayu Xuexizhe Kouyu Yuliaoku* (华语学习者口语语料库) [Chinese Learners Spoken Corpus]. https://corpus.eng.ntnu.edu.tw/cqpweb/chinese_learners_sp/

19. National Taiwan Normal University. (2017). *Huayu Xuexizhe Shumian Yuliaoku* (华语学习者书面语料库) [Chinese Learners Written Corpus]. https://corpus.eng.ntnu.edu.tw/cqpweb/chinese_learners_wr/
20. National Taiwan Normal University. (s.d.). *Zhongwen Xueshu Yuliaoku* (中文学术语料库) [Chinese Academic Corpus]. <https://corpus.eng.ntnu.edu.tw/cqpweb/>
21. National Taiwan Normal University. (s.d.). *Zhongwen Minfa Yuliaoku* (中文民法语料库) [Civil law corpus]. <https://corpus.eng.ntnu.edu.tw/cqpweb/>
22. National Taiwan Normal University. (s.d.). *Zhongwen Wanglu Shangwu Yuliaoku* (中文网路商务语料库) [Business Chinese Corpus]. <https://corpus.eng.ntnu.edu.tw/cqpweb/>
23. National Taiwan Normal University. (s.d.). *Zhongwen Wanglu Kejilei Yuliaoku* (中文网路科技类语料库) [Technology Chinese Corpus]. <https://corpus.eng.ntnu.edu.tw/cqpweb/>
24. National Taiwan Normal University. (s.d.). *Zhongwen Wanglu Lüyou Yuliaoku* (中文网路旅游语料库) [Tourism Chinese Corpus]. <https://corpus.eng.ntnu.edu.tw/cqpweb/>
25. Peking University. Zhan, W. (詹卫东), Guo, R. (郭锐), y Chen, Y. (谌贻荣). (s.d.). *Beijing Daxue Xiandai Hanyu Yuliaoku* (北京大学现代汉语语料库) [The CCL Corpus of Chinese Texts]. http://ccl.pku.edu.cn:8080/ccl_corpus
26. Shanghai Jiaotong University. (2011). *Yixue Ying-Han Shuangyu Pingxing Yuliaoku* (医学英汉双语平行语料库) [English-Chinese Medical Parallel Corpus (ECMPC)].
27. Shaoxing University. (2009). *Zhongguo Falü Fagui Han-Ying Pingxing Yuliaoku* (中国法律法规汉英平行语料库) [Parallel Corpus of China's Legal Documents].
28. Shaoxing University. (2010). *Honglouloumeng Han-Ying Pingxing Yuliaoku* (《红楼梦》汉英平行语料库) [Parallel Corpus Chinese-English of *A Dream of Red Mansions*].
29. University of California Los Angeles. Tao, H. (陶红印), y Xiao, R. (肖忠华). (2012). *The UCLA Corpus of Written Chinese (2nd edition)*. <http://114.251.154.212/cqp/ucla2/>
30. Universidad Autónoma de Madrid. (2011). *Corpus CORAL CHINA*. <http://cartago.illf.uam.es/dat/c-oral-chino>
31. Universidad de Valladolid, Universidad de Santiago de Compostela, Universidad de Barcelona, y Universidad de Vigo. (2025). *Parallel Corpus Chinese<>Spanish, PaCheS*. https://www.pacores.eu/los-corpus/#corpus_paches