ISSN 2007-5308

Compilación de un corpus de aprendientes chinos de español

Compilation of a learners' Corpus of Chinese Students learning Spanish

DOI: 10.32870/mycp.v13i39.898

Rodrigo Muñoz Cabrera[1]
Lan Zhang[2]

Resumen

El análisis de la producción de los aprendientes de una lengua extranjera resulta vital para definir los parámetros por los que ésta se adquiere. Es por ello que resulta indispensable la elaboración y utilización de *corpus* de aprendientes. La pandemia de covid-19 provocó cambios en la metodología docente. Durante la misma, debido a la imposibilidad de impartir las clases de manera presencial y al estar el alumno obligado a enviar sus trabajos en formato electrónico para su corrección, fue posible la compilación de un *corpus* empleando los trabajos de aprendientes chinos que estudian español en la Universidad de Nankai (China). El objetivo de este artículo es detallar los pasos para la creación de un *corpus* textual, su compilación y las herramientas a emplear para su análisis, específicamente el programa AntConc. Después de la creación y análisis del *corpus,* se concluye que utilizando el *corpus* se pueden identificar las carencias y dificultades a las que se enfrentan los estudiantes, lo que permite conformar modelos de aprendizaje lingüístico y mejorar los ya existentes. Además, su incorporación en el aula permite individualizar el aprendizaje lingüístico del alumnado y generar teorías sobre el uso que se hace del lenguaje.

Palabras clave: análisis de errores, español como segunda lengua, *interlengua*, lingüística de *corpus,* transferencia lingüística.

Abstract

The analysis of the production of learners of a foreign language is crucial for defining the parameters by which it is acquired. Therefore, the development and use of learner corpora are indispensable. The Covid-19 pandemic brought about changes in teaching methodology. During this period, due to the impossibility of conducting in-person classes and the necessity for students to submit their work electronically for correction, it became possible to compile a corpus using the essays of Chinese learners studying Spanish at Nankai University (China). This article aims to detail the steps for creating a textual corpus, its compilation, and the tools to be used for its analysis, specifically the program AntConc. After the creation and analysis of the corpus, it is concluded that, by using the corpus, one can identify the deficiencies and difficulties faced by students, which allows for the formation of linguistic learning models and the improvement of existing ones. Moreover, its incorporation in the classroom enables the individualization of students' linguistic learning and the generation of theories about language use.

Keywords: corpus linguistics, error analysis, interlanguage, linguistic transfer, Spanish as a second language.

Artículo recibido el 7 de enero de 2024 y dictaminado el 2 de mayo de 2024.

1. Universidad Internacional de La Rioja (UNIR). Avenida de la Paz, 137. 26006-Logroño, España. ORCID: https://orcid.org/0000-0001-7890-4083 Correo electrónico: rodrigo.munozcabrera@unir.net
2. Universidad Complutense de Madrid. Av. Complutense, s/n, 28040-Madrid, España. ORCID: https://orcid.org/0000-0002-4504-1004 Correo electrónico: lanzhang@ucm.es

1. Introducción

Este artículo tiene como propósito detallar los pasos para la creación de un *corpus* textual con el que se puedan identificar los errores (en particular, los incurridos a causa de imprecisiones gramaticales y léxicas, transferencias, interferencias y calcos lingüísticos, y fallos de naturaleza intralingüística) cometidos por aprendientes chinos que estudian español en la Universidad de Nankai (China). La pandemia nos permitió acometer esta empresa debido (entre otros factores) a la imposibilidad de impartir las clases de manera presencial y a estar el alumno obligado a enviar sus trabajos en formato electrónico para su corrección. Con dicho repertorio textual nos fue posible llevar a cabo un proyecto iniciado hace ya siete años, que esperamos sirva como base para futuros estudios encaminados a la mejora del nivel de español entre nuestros aprendientes *sinohablantes*. No obstante, esta metodología puede ayudar a cualquier profesor de español como lengua extranjera o ELE (independientemente de la nacionalidad de su alumnado) a la hora de analizar y corregir los errores que cometen sus discentes.

La configuración de un *corpus* textual implica, por un lado, la necesidad de disponer de textos para su observación y, por otro, el uso de herramientas de análisis de *corpus*. Los escritos albergan una serie de evidencias que, mediante un nuevo marco metodológico de observación, nos será posible identificar y evaluar (Tognini-Bonelli, 2001, p. 3). A ello cabe añadir los postulados teóricos de dicha disciplina, los cuales nos servirán para analizar los datos extraídos del *corpus* y formular las hipótesis que de ellos se deriven. En aquella época no tan lejana de restricciones, cuando un porcentaje nada desdeñable del profesorado se vio forzado a llevar a cabo su actividad docente en línea, tuvimos la oportunidad de aumentar exponencialmente nuestro *corpus* de aprendientes de español. No se trata de un aspecto banal: el aprendiz chino tiende a realizar los deberes a mano y, en menor medida, en computador. Como norma general, éstos entregan sus tareas de manera presencial a la llegada al aula y, en muy pocas ocasiones, mediante correo electrónico o por medio de (a modo de servicio de mensajería) la plataforma WeChat.

2. Marco teórico

2.1. Lingüística de corpus y corpus

Como ya hemos detallado, nuestro propósito ha sido compilar un *corpus* textual con el cual verificar la evolución del estudiantado y apreciar qué aspectos lingüísticos y gramaticales eran en los que erraban con mayor frecuencia. En primer lugar, y una vez detallados nuestros objetivos, estimamos conveniente describir qué es la lingüística de *corpus*. Se trata del estudio de un gran volumen de datos lingüísticos; es el análisis asistido por computadora de una gran recopilación de manifestaciones lingüísticas en textos, transcripciones o grabaciones (McEnery & Hardie, 2012, p. 1). Gracias a esta metodología, ya es posible estudiar el lenguaje bajo un método de base científica (Leech, 1992, p. 112).

Se trata de un análisis de *lenguaje real*, escrito por personas (no creado artificialmente o *ex profeso)*, que no ha sido fruto de la intuición o de la introspección del investigador. Mediante la lingüística de *corpus* se lleva a cabo un estudio del lenguaje de manera científica, en donde es posible examinar un alto volumen de datos. Asimismo, la informática ha dotado a la lingüística con herramientas para observar el lenguaje producido de manera natural, capaz de procesar datos textuales con rapidez y eficacia (McEnery & Wilson, 2001, p. 17).

Inicialmente, la lingüística de *corpus* influenció sobremanera la enseñanza del inglés como lengua extranjera (*English Language Teaching* o ELT) para, seguidamente, ser empleada en estudios sobre el aprendizaje de segundas lenguas en general (Ellis, 2003), práctica de metodologías en el aula e incluso material de examen (Millán, 2006). En lo que se refiere a la mejora de la adquisición del inglés como lengua extranjera, Harper Collins publicó en 1987 (bajo la dirección de John Sinclair) el primer diccionario en el que se empleó un *corpus* lingüístico para así facilitar el aprendizaje de la lengua inglesa a los aprendientes. En su compilación se usaron textos *reales* y *auténticos*, en el que se incluían no sólo definiciones, sino también ejemplos en contexto. Dicha obra fue bautizada como *Cobuild English Language Dictionary* (Baker et al., 2006, p. 108), repertorio de incalculable valor al proporcionar al aprendiente innumerables ejemplos en contexto extraídos de textos redactados por anglófonos. Posteriormente, editoriales como Longman o Cambridge University Press han conformado sus propios *corpus* de aprendizaje que se emplean regularmente para crear (entre otros) materiales didácticos de enseñanza del idioma inglés.

Ya hemos mencionado que la lingüística de *corpus* estudia el lenguaje y esos ejemplos de su uso (independientemente del formato en el que se hayan compilado) se almacenan en lo que se denomina *corpus*. Podemos definir el concepto de *corpus* como una recopilación de textos (escritos, orales o transcritos) cuya cantidad y contenido se consideran *representativos* de una determinada lengua con el propósito de realizar análisis lingüísticos de variada índole (Tognini-Bonelli, 2001, p. 53). En el caso que nos ocupa, nos referimos a un *corpus* compilado con la participación de aprendientes que estudian una lengua extranjera (en este caso, el español). En términos generales, éste se nutre de tareas escritas sobre temas previamente acordados en clase (Baker et al., 2006, p. 103). Este tipo de recopilaciones textuales se denomina *corpus de aprendientes*, a saber: "textos concebidos para el estudio de una segunda lengua no escritos por hablantes nativos, sino redactados por los propios alumnos" (James, 1992, p. 190). Los *corpus* de aprendientes se emplean (en mayor medida) para individualizar el aprendizaje lingüístico del alumnado y generar teorías sobre el uso que se hace del lenguaje. En la actualidad, este tipo de *corpus* también se emplea en la detección y corrección de errores (Gamon et al., 2013), la identificación de lenguas maternas (Tetreault et al., 2013) y en la formulación de hipótesis sobre trasferencias lingüísticas (Swanson & Charniak, 2014), entre otros campos.

Nuestro propósito es emplear nuestro *corpus* en la observación de la adquisición de segundas lenguas (*Second Language Acquisition* o SLA por sus siglas en inglés), con el fin de conformar modelos de aprendizaje lingüístico y así suplementar o mejorar los ya existentes. En consecuencia, el análisis de la producción de los aprendientes resulta vital para definir los parámetros por los que se adquiere una lengua extranjera. Mediante la pormenorización de los errores cometidos es factible discernir las estrategias que éstos adoptan al capacitarse en el uso del español como lengua extranjera (Castillejos, 2009, p. 675). Con ello se posibilita la formulación de estrategias para solventar los escollos gramaticales que afectan a la mayoría de los aprendientes a la hora de comunicarse en español. De igual manera, el estudio de lenguaje natural (creado por los propios aprendientes) resulta muy útil a la hora de identificar su *interlengua* (Wang et al., 2015, p. 119).

2.2. Análisis de errores

En lo que respecta a la enseñanza de una segunda lengua o L2, el análisis de errores cumple un papel preponderante, puesto que no sólo permite conocer las lagunas gramaticales de que adolece el aprendiente, sino que también facilita la elaboración de nuevas propuestas para subsanar tales deficiencias. Una nueva tendencia metodológica hace hincapié en configurar destrezas comunicativas, en las que los errores no son considerados de gran importancia, siempre y cuando no dificulten la comunicación (Castillejos, 2009, p. 677). No obstante, la identificación de los errores, su estudio y la propuesta de estrategias para paliarlos y obtener un conocimiento y un empleo de la gramática más eficaces no deja de ser primordial en el aprendizaje. De igual modo, acometer dicha tarea durante los diferentes estadios por los que un alumno transcurre durante su periodo de aprendizaje puede contribuir a la comprensión de las estrategias que éstos emplean a la hora de poner en práctica lo aprendido de una segunda lengua (Ellis, 2003).

Llegados a este punto, estimamos necesario destacar la vaguedad en lo que denominamos *error,* un aspecto de difícil consenso ya que entraña la concepción personal que cada investigador posee sobre esta noción, puesto que lo que para unos se trata de un fallo, para otros puede no serlo necesariamente. En consecuencia, la indeterminación ejerce un papel fundamental a la hora de proponer solución a los errores analizados (Castillejos, 2009, p. 687). A modo de ejemplo, es complicado llegar a un acuerdo a la hora de catalogar como desaciertos los de naturaleza gramatical y morfológica. Si a ello le añadimos que la *indeterminación* puede aparecer en la corrección que se propone, hemos de considerar que la consideración de *error* es un tema que suscita divergencia (James, 1998, p. 11).

2.3. Interlengua

Selinker y Rutherford (2013) proponen otra perspectiva en el análisis de los errores, a la que denominan *interlengua*, que resulta de la interacción de dos sistemas lingüísticos (el de la lengua materna, o L1, y el de la L2 que se aprende), una interacción que tiene como resultado la creación de una tercera lengua, un sistema lingüístico independiente. Selinker define la *interlengua* como "reglas y pautas lingüísticas que el alumno de una segunda lengua crea para uso personal" (1974, p. 31). Este tipo de errores que cometen los

discentes pueden ser motivados por el fenómeno de la transferencia (como consecuencia de la influencia ejercida por la lengua materna) o por causas intralingüísticas (producidas por influencia de la lengua meta). Mientras los errores motivados por la transferencia lingüística son producidos por la estructura de la L1, los de entre lenguas no tienen origen en ésta, sino en la L2 (Castillejos, 2009, p. 683). Los errores producidos entre pares de idiomas no son siempre el resultado de contrastar la lengua materna con la segunda lengua, sino que también están relacionadas con una interpretación específica de la L2. Se trata de un fenómeno universal que aparece en todo proceso de aprendizaje lingüístico (Els, 1984, p. 21). En términos generales, la *interlengua* es una lengua distinta y su observación se lleva a cabo mediante el empleo de los *corpus* de aprendientes.

El estudio de los errores nos permite obtener información de cómo se aprende un idioma y refleja la creación de estrategias por parte del alumnado para su aprendizaje, además de proporcionar la información lingüística que un alumno adquiere del idioma objeto de estudio en un momento determinado. Consideramos asimismo necesaria la identificación de ciertas estrategias empleadas, como pueden ser la omisión en el uso de estructuras gramaticales complejas por parte del aprendiente. La evaluación de la competencia adquirida es un válido indicador de la evolución en el aprendizaje lingüístico (Yip, 1995, p. 5).

Selinker y Rutherford (2013) distinguen cinco factores determinantes en la adquisición de una segunda lengua, a saber:

1. Transferencia lingüística (resultado de la interacción de la lengua materna del aprendiente).
2. Transferencia educativa (resultado condicionado por el tipo de instrucción que se ha recibido al aprender una L2).
3. Estrategias de aprendizaje de una segunda lengua (producidas por la asociación que el aprendiente crea a través del material didáctico empleado en el aprendizaje).
4. Estrategias comunicativas en una segunda lengua (su producción se obtiene mediante la asociación que el aprendiente crea mediante la comunicación con hablantes nativos de la lengua objeto de estudio).
5. Generalización en exceso del material lingüístico de la lengua objeto de estudio (resultado de llegar a conclusiones excesivamente genéricas de las reglas semánticas y sintácticas de la L2).

Son, en gran medida, esas transferencias, estrategias y generalizaciones las que provocan los errores por parte del estudiantado.

2.4. Traducción

En otro orden de cosas, cabe destacar el amplio uso que los aprendientes chinos hacen de la traducción. Resulta evidente que copiar patrones gramaticales de la L1 a una segunda lengua no es una práctica acertada. Sin embargo, la traslación forma parte del aprendizaje y ayuda al alumnado a verter, no sólo palabras, sino ideas y conceptos a una lengua meta de manera más adecuada. En definitiva, se trata de concebir la traducción como "un acto de comunicación y no como una mera actividad de trasladación lingüística" (Cerezo, 2019, p. 240). Además, la traducción no se limita a verter un texto de una lengua origen a una lengua meta, se trata de una disciplina que puede ayudar en gran medida a la hora de adquirir un nuevo idioma. Como menciona Hurtado: "El ámbito de la traductología es vastísimo, introduciéndose en campos que necesitan de muchas especialidades" (1999, p. 10). Siguiendo esta premisa, estimamos que los estudios de traducción no sólo están enfocados en la mera traslación y en el examen de los lenguajes de especialidad, sino que pueden esgrimirse como una herramienta valiosa empleada (entre otras cosas) para estudiar una L2 desde un enfoque contrastivo con la L1 (Cerezo, 2014, p. 285).

En consecuencia, y contemplado desde esta perspectiva, estimamos que la traslación no siempre es un método equívoco a adoptar en el aprendizaje de un segundo idioma. Su utilización queda patente en los niveles iniciales del aprendizaje, cuando la necesidad de comunicación sobrepasa ampliamente los recursos expresivos adquiridos y la traducción de términos de la lengua materna es un recurso muy frecuente (Loffler, 2011, p. 24). Debido a que el aprendiz *sinohablante* la emplea (aunque en demasía), el debate versaría sobre la adecuación o no de introducir nociones de traslación en el aula para una mayor clarificación al respecto. Sería una opción a la hora de poner en evidencia ciertas diferencias entre la lengua materna y la segunda lengua por medio de la gramática comparada y de la traducción libre (en contraposición a la literal), para hacer entender a los aprendientes que las representaciones semánticas en dos idiomas requieren (en mayor medida) de estructuras gramaticales, sintácticas y léxicas diferentes.

3. Compilación y uso del *corpus*

Inicialmente, y como ocurre en estos casos, el porcentaje de aprendientes dispuesto a participar en este estudio fluctuaba entre los diversos grupos. No obstante, tras casi tres años de docencia en línea por motivos ampliamente conocidos (y debido a la imperiosa necesidad de recibir y enviar los trabajos por vía telemática), el número de escritos incluidos en nuestro *corpus* experimentó un incremento sustancial, con tareas de aprendientes de tres nuevas promociones. Una vez explicado el uso que íbamos a hacer de los trabajos, ningún aprendiente puso objeciones a la hora de emplear sus tareas para conformar nuestro *corpus*. La compilación textual realizada durante siete años ha sido bautizada con el nombre de *Corpus de español de estudiantes chinos* (CEEC) y la componen 504 archivos textuales que albergan un total de 322,434 voces. Dicho *corpus* fue elaborado en la Universidad de Nankai (Tianjin, China) durante 2016 y 2023.

3.1. Parámetros

Una vez enumerados nuestros objetivos, procedemos a detallar las herramientas y metodología empleadas. A la hora de delimitar la investigación, y debido al aumento cualitativo que experimenta el alumnado durante los cuatro años que dura el Grado en Filología Hispánica, estimamos adecuado establecer previamente los parámetros que definirán nuestro trabajo. El primero de ellos fue seleccionar a los estudiantes de segundo año como objeto de nuestra investigación, al considerar que son ellos quienes muestran el mayor número de errores característicos sujetos a lo nuestro (los provocados por emplear estructuras gramaticales y expresiones del mandarín, ciertos aspectos culturales, calcos procedentes de otras lenguas y el empleo de la traducción literal, entre otros). Los alumnos, una vez concluido el primer año de estudio, adquieren las competencias básicas para expresarse por escrito en castellano y es en el segundo curso donde se inicia el verdadero reto de escribir con claridad en dicha lengua, aunque con evidentes limitaciones.

El segundo parámetro adoptado se encuentra relacionado con la temática de los trabajos. La mayoría de ellos son ejercicios por los que deben demostrar que han aprendido o entendido las destrezas (gramaticales, léxicas, culturales, históricas, etc.) explicadas en el aula. Del mismo modo, se incluyen trabajos finales de asignaturas de historia y de comprensión lectora, al no existir gran-

des divergencias entre estos dos tipos de obras, con excepción de una mayor pulcritud y extensión en estas últimas. Los tipos textuales de los trabajos son descriptivo, narrativo, expositivo y argumentativo, siendo los dos primeros los que más abundan en nuestro repertorio.

En lo que respecta al tercer factor (el tamaño), el número de palabras que los componen varía en 150 y 1,000. Desde un punto de vista lingüístico, no resulta muy coherente elegir muestras textuales del mismo tamaño o extensiones similares. Inicialmente, los *corpus* se compilaban siguiendo esa premisa, la cual es difícilmente justificable hoy en día. La integridad y la representatividad de textos completos posee una mayor relevancia, que la dificultad que entraña conciliar textos de diferentes dimensiones (Sinclair, 2005, pp. 6-7). Siguiendo esta premisa, en el *corpus* sólo encontramos textos completos, independientemente del número de palabras que contengan.

El cuarto aspecto trata de la cantidad. En este caso, podemos afirmar que la máxima *cuanto más grande, mejor,* sí se adecua a las necesidades de configuración de este *corpus*. En términos generales, los datos que se extraen de un *corpus* de pequeñas dimensiones, pero bien diseñado, serán más y mejores que los procedentes de uno mayor que no ha sido creado conforme a nuestras necesidades (Bowker & Pearson, 2002). No obstante, ello no quiere decir que no sea aconsejable compilar un *corpus* de gran tamaño; todo depende de la *calidad* de los textos. Por tanto, cuantos más textos satisfagan las exigencias que configuran nuestra investigación, mejor, pero siempre es imprescindible otorgarle más valor a la *calidad* que a la *cantidad* (McEnery & Hardie, 2012). Por otro lado, nuestro propósito es que sea un *corpus sincrónico*, en donde los escritos hayan sido redactados en el mismo espacio temporal o en circunstancias análogas; una instantánea que permita estudiar el uso del lenguaje en un momento determinado (Baker et al., 2006, p. 153). Ello nos permitirá examinar los problemas lingüísticos y culturales de estos aprendientes en un periodo específico de su instrucción (el segundo año en el Grado de Filología Hispánica), tratándose asimismo de un *corpus dinámico* para añadir más archivos de alumnos de segundo año de otras promociones y, de ese modo, apreciar el mayor número posible de faltas con las que evaluar, con una mayor claridad, las necesidades específicas del alumnado chino al aprender el idioma español.

3.2. Conversión de textos

Se trata de una idea que fue concebida a inicios de 2016, momento en el que solicitamos a los aprendices de segundo curso (a título personal y de manera voluntaria) la entrega de las tareas redactadas en el programa Microsoft Word en formato doc o docx. Una vez recibidas, éstas fueron modificadas y transformadas a texto sin formato o txt: a saber, documentos que sólo contienen palabras y, posiblemente, anotaciones (Weisser, 2016, p. 36). Finalizada la conversión, es aconsejable revisar los textos (por medio de Microsoft Word, WordPad o Block de notas) para comprobar que el cambio sí se ha llevado a cabo con éxito. Observaremos que gráficos, dibujos y fotos (si hubiere) habrán sido eliminados, que el interlineado es de 1.0 a la izquierda, que el tipo de fuente es Courier New y el tamaño ha pasado a 10.5. De igual modo, es importante comprobar que se mantiene un solo espacio entre palabras. Se trata de un aspecto muy importante a tener en cuenta entre el profesorado, ya que el uso de diferentes tipos de letra provoca errores al enviar ficheros y visualizarlos en otra configuración o sistema operativo. En muchas ocasiones, los aprendientes chinos emplean el mismo tipo de letra que usan para escribir en chino mandarín (principalmente DengXian y SimSum) a la hora de hacerlo en español, hecho que provoca anomalías, puesto que el espacio entre letras y palabras aumenta. Todo ello es clave si nuestro deseo es revisar adecuadamente los escritos con el *software* que emplearemos para realizar el examen lingüístico.

3.3. Herramienta de análisis de corpus: AntConc

En los últimos 50 años el computador personal ha visto reducido su precio e incrementado su velocidad, con el que se obtiene una alta calidad en el procesamiento de textos. En la actualidad es posible ejecutar el tipo de análisis de *corpus* en cuestión de minutos con una computadora. El computador en sí ha dejado invalidadas las críticas que etiquetaban a la lingüística de *corpus* como un *pseudoprocedimiento*, en particular en lo que se refiere a la extracción de formas nominales en grandes recopilaciones textuales (McEnery & Wilson, 2001, p. 17).

Un valor añadido a la rápida evolución de la informática recae en la aparición de programas que satisfagan nuestras necesidades. A la hora de seleccionar una herramienta de análisis de *corpus,* nuestro propósito fue em-

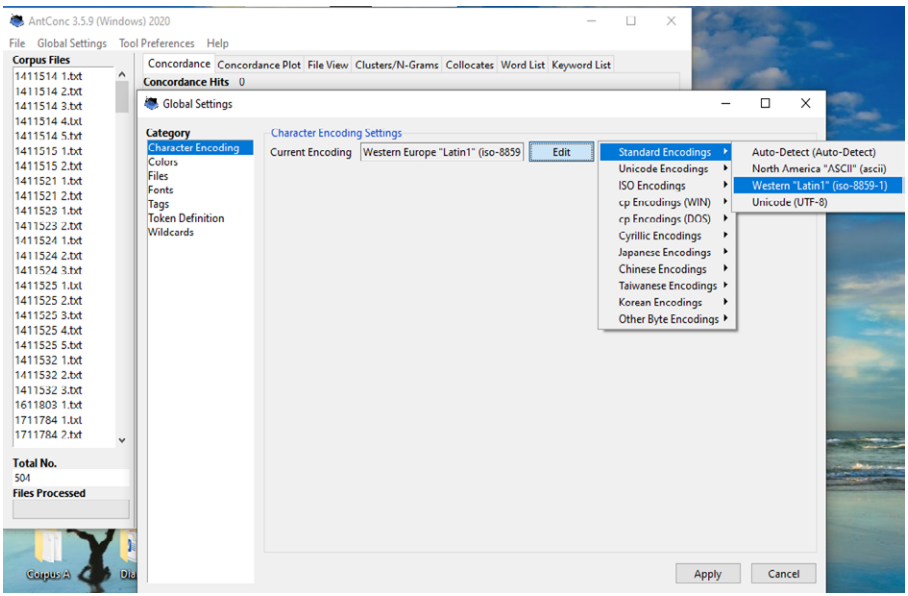
plear un programa gratuito, de fácil uso y dotado de las aplicaciones básicas, necesarias para la examinación textual. AntConc es un *freeware* de análisis de *corpus,* diseñado específicamente para su uso en el aula. Permite la búsqueda de concordancias, vocablos, palabras clave, búsqueda de asociación de voces y ubicación de palabras en contexto, además de otras opciones (Anthony, 2004, p. 7). Se trata de un programa autoinstalable, disponible sin costo alguno y sin necesidad de registrarse. Lamentablemente, el escollo radica en que este *software* está únicamente disponible en inglés. A continuación procedemos a detallar las funciones básicas del programa AntConc, versión 3.5.9 (Windows) 2020, la más actualizada y la empleada para nuestra investigación.[3]

3.4. Configuración preliminar

Antes de comenzar el uso del programa, es necesario llevar a cabo una serie de ajustes para adecuarlo a nuestras necesidades. Al abrir AntConc, debemos presionar sobre *File* (arriba a la izquierda) y cargar los archivos que compongan nuestro *corpus*. Tras ello, es necesario entrar en *Global Settings* (pestaña situada a la derecha de *File)* y seleccionar el sistema de codificación que, por defecto, es Unicode (UTF-8). Ese sistema no es el adecuado para estudiar las lenguas romances ya que, entre otras cosas, no detecta las tildes. Por ello, es imprescindible sustituirlo por Western "Latin 1" (iso -8859-1) y, tras ello, pulsar *Apply* (abajo). No se trata de una banalidad, puesto que (entre otras cosas) los signos iniciales de interrogación y de exclamación serán interpretados por el *software* como un error, lo que provocará cierta distorsión en los resultados. Tras este sencillo ajuste, podremos iniciar nuestra tarea (véase figura1).

3. Para descargar el programa AntConc, versión 3.5.9 (Windows), al igual que los tutoriales en inglés e instrucciones de uso en español, diríjase a https://www.laurenceanthony.net/software/antconc/. Si desea visualizar videos en español sobre su empleo, acceda al siguiente enlace de Youtube: https://www.youtube.com/results?search_query=antconc+tutorial+espa%C3%B1ol

Figura 1
Configuración de AntConc 3.5.9. (Windows) 2020



Fuente: elaboración propia

Otro dato a destacar es que, a la izquierda de la imagen aparecen los textos que conforman el *corpus,* en la ventana de nombre *Corpus Files;* en este caso son 504 en total.

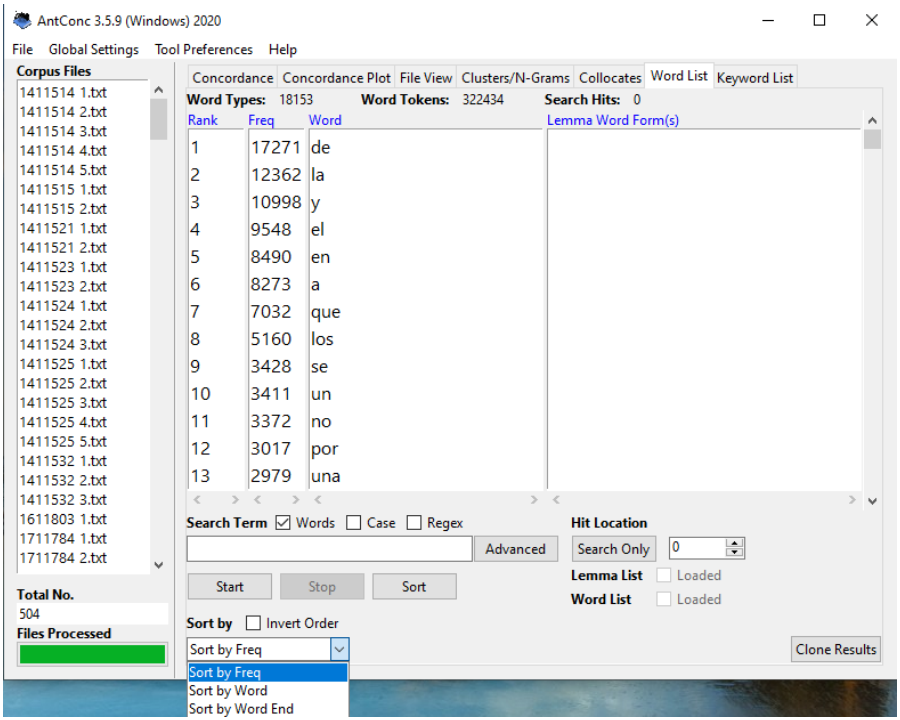
3.5. Listado de palabras

La primera información a extraer de un *corpus* es descubrir las voces que éste alberga y su número. En la parte superior de la pantalla aparecen siete pestañas *(Concordance, Concordance Plot, File View, Clusters/N-Grams, Collocates, Word List y Keyword List)*. Debe darse clic sobre *Word List*. Tras ello, basta pulsar el botón *Start* (en la parte inferior izquierda) para generar un listado de palabras y descubrir las veces que se repiten en orden alfabético, por su frecuencia de aparición e incluso por medio de con qué letra(s) termina cada vocablo. Ello nos permitirá visualizar los errores o las voces (y su lematización, si se desea), pudiendo también identificar aquellas que se han escrito de manera incorrecta. De igual modo, y si se quiere evitar el recuento de palabras vacías (artículos,

preposiciones, adverbios, etc.), es posible añadir un archivo con ese tipo de voces *(Stop List)* para que no sean contabilizadas. El recuento de palabras es el primer paso en el que podremos apreciar qué aspectos lingüísticos son dignos de observación. A modo de ejemplo, dicha lista puede mostrarnos voces cuyo porcentaje de aparición sea inusualmente alto o bajo e investigar qué motivos provocan tal anomalía.

En este caso puede comprobarse que nuestro *corpus* alberga 18,153 palabras diferentes *(Word Types)* y un número total de 322,434 voces *(Word Tokens)*. En los términos de búsqueda (*Search Term*, abajo a la izquierda y en negrita) se seleccionó la búsqueda de palabras *(Words)*, se optó por la disposición por frecuencia de aparición *(Sort by Freq)*, aunque es posible un ordenamiento alfabético, por terminación de palabra o invertir sus órdenes (véase figura 2).

Figura 2
Listado de palabras o *Word List*



Fuente: elaboración propia

En la figura 2 podemos comprobar —como era de suponer— que en los primeros puestos aparecen artículos, pronombres, preposiciones, etc., mostrándose en la columna a su izquierda *(Freq)* el número de veces que se repiten. Como ya hemos comentado, es posible incluir una lista de palabras vacías *(Stop Word List)* clicando en *Tool Preferences* (en la parte superior, a la izquierda), ir a la categoría *Word List* y almacenar un archivo que contenga tales palabras si es nuestro deseo enfocar nuestro estudio en verbos, sustantivos o adjetivos.

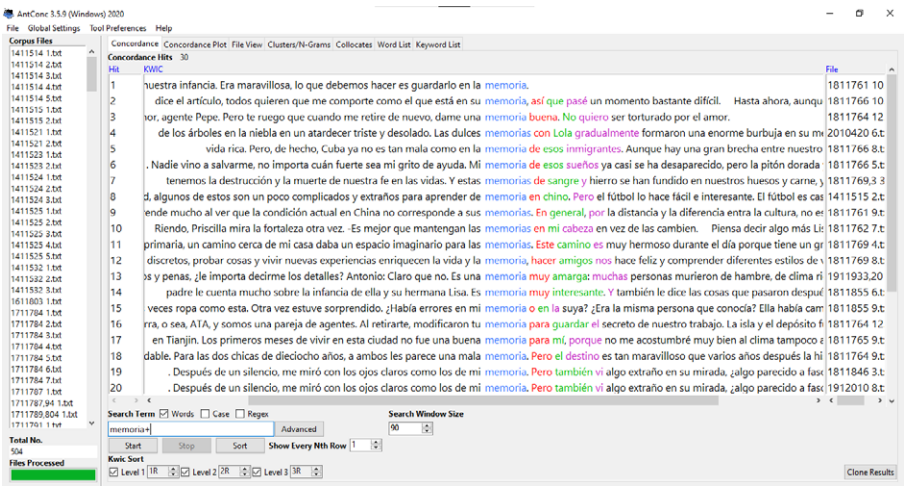
3.6. Búsqueda de concordancias

La herramienta para búsqueda de concordancias *(Concordancer)* ha demostrado ser una ayuda altamente eficaz durante la adquisición de una segunda lengua, puesto que facilita el estudio de la gramática, aprendizaje de vocabulario, *colocaciones* y estilo (Wang et al., 2015).

En la figura 3 podemos visualizar el aspecto del programa al emplear la herramienta de concordancia (véase en la parte superior, a la izquierda, *Concordance*, resaltada en color blanco). Las siete pestañas están diseñadas para que, durante su empleo, sea posible acceder a la mayoría de las aplicaciones básicas, sin necesidad de cambiar de pantalla. Una característica primordial en el diseño de un buen programa es evitar menús desplegables y ventanas adicionales, puesto que tienden a confundir al usuario. Obsérvese que teclados virtuales, listados, datos y ajustes de ventana son similares a los que existen en los programas informáticos de mayor difusión en el mercado.

La búsqueda de concordancias posee un gran número de características que hacen que sea una herramienta eficaz, no sólo para aprendientes sino también para profesores e investigadores (Anthony, 2004, p. 8). Comencemos con una búsqueda simple de palabras. A modo de ejemplo, escribimos *memoria+*, en el casillero de búsqueda de palabras/términos (*Search Term*, abajo) y pulsamos en *Start*. El comodín + nos permitirá buscar la palabra *memoria*, además de ésta con una letra más al final, ya que deseamos examinar también su forma plural. En las búsquedas se puede hacer uso de una serie de comodines *(Wildcards)* y añadirlos a una palabra/lema, los cuales se pueden visualizar en *Global Settings*, arriba a la izquierda.

Figura 3
Resultados de la búsqueda *memoria+*



Fuente: elaboración propia

Tal y como aparece en la parte superior izquierda (*Concordance Hits*, en negrita), constatamos que han aparecido 30 resultados, tanto en forma singular como en plural. Justo encima de donde hemos escrito el vocablo objeto de búsqueda, visualizamos las opciones (a la derecha de *Search Term*, en negrita) *Words*, *Case* y *Regex*. Al seleccionar *Words*, el programa buscará una palabra o término, conjuntamente con los comodines que hayamos utilizado. Si optamos por la casilla *Case*, se seleccionará la palabra de búsqueda tal y como haya sido escrita: en minúscula, mayúscula inicial o todo en mayúscula. Por defecto, aparecerán todas las formas escritas de la voz sujeta a estudio, por lo que se recomienda optar por dicha opción si se desea diferenciar entre mayúscula y minúscula. *Regex* permite realizar pesquisas por medio de expresiones y no por palabras. A la derecha de estas casillas se observa la pestaña *Search Window Size*. Con ella, se amplía o disminuye la línea en donde se encuentra ubicada la palabra clave elegida (en contexto) que aparece en pantalla. Somos de la opinión de que cuanto más texto veamos, mejor, y por ello elegimos visualizar hasta 90 espacios.

En el centro de la pantalla central aparece la palabra a analizar en azul claro. Nótese que la primera, segunda y tercera voces a su derecha muestran colores diferentes. Ello se debe a que en *Kwic Sort* (en la parte inferior de la

imagen, a la izquierda) existen tres niveles diferentes *(Level 1, Level 2 y Level 3)*, en donde se ha decidido destacar la primera, segunda y tercera palabras a la derecha del vocablo a investigar (1R, 2R y 3R). Se puede elegir hasta la vigésima palabra a la izquierda o a la derecha de la palabra de búsqueda, aunque lo más común es seleccionar las tres primeras que figuran a su derecha o a su izquierda (en ese caso, 1L, 2L y 3L). También es posible descartar dicha opción. Si así se decide, únicamente la palabra de búsqueda se visualizará en azul; el resto aparecerá en negro.

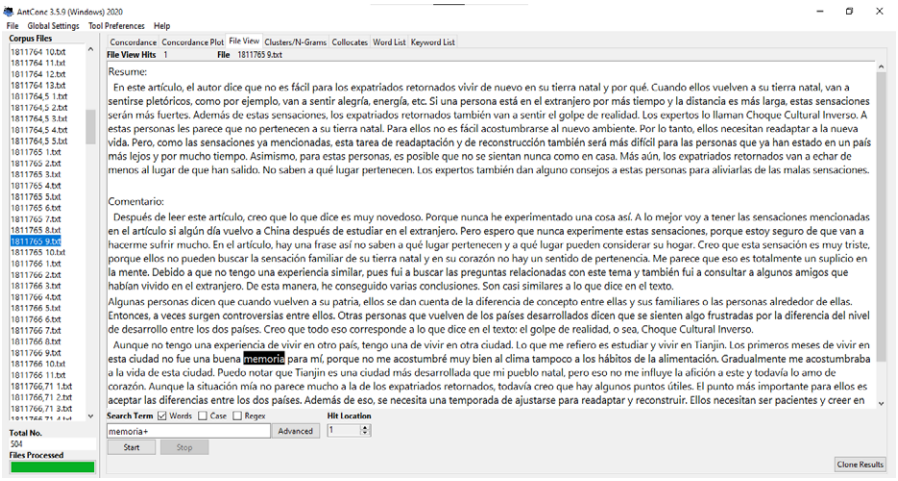
Los ejemplos se encuentran numerados en la columna de la izquierda *(Hits)* y el nombre del fichero en donde se encuentra cada uno figura en la columna de la derecha *(File)*. Se trata de un dato clarificador, puesto que si los resultados de la búsqueda aparecen en diferentes archivos, escritos por distintos aprendientes, implica que se trata de un hecho digno de análisis. Si, por el contrario, éstos se encuentran en un mismo archivo o en diferentes textos redactados por un mismo alumno, puede tratarse de un error puntual producido por un aprendiente en particular que no puede extrapolarse a toda la comunidad estudiantil sujeta a estudio.

Al dar clic sobre cualquiera de los ejemplos alineados en azul claro, AntConc nos trasladará, automáticamente, al archivo textual donde se encuentra el ejemplo seleccionado y nos lo mostrará con todo su contexto en la ventana *File View*, arriba. A modo de ejemplo, mostramos en la figura 4 lo que aparece al presionar en la decimoséptima concordancia *(Hit)*. Comprobaremos que la imagen que se nos muestra no es la de *Concordance*, sino la de *File View* (véanse de nuevo las pestañas situadas en la parte superior, sobre fondo en blanco), la visualización del texto elegido. En esta ocasión es posible apreciar la voz (sobreimpresionada en negro) con todo su contexto. En esta pantalla es posible comprobar el archivo textual en su totalidad, presionando la pestaña alargada situada en el margen derecho, deslizándola hacia arriba o hacia abajo con el ratón. A la izquierda, en la columna *Corpus Files*, apreciaremos que el nombre del archivo inspeccionado aparece resaltado en azul.

En este caso constatamos la confusión entre los estudiantes a la hora de emplear *memoria*, *recuerdo* y *experiencia*. El sustantivo *recuerdo* o el verbo *recordar* se traducen en mandarín como *huíyì* (回忆), mientras que memoria es *jìyì* (记忆) y, para hablar sobre una experiencia personal, se emplea *jīnglì* (经历). En la frase seleccionada, "Los primeros meses de vivir en esta ciudad no fue una buena memoria para mí", comprobamos un uso incorrecto en el sustantivo *memoria*. Tras consultar con nativos, éstos aducen que no usarían

jìyì al traducir esa frase en mandarín, sino *huíyì* (recuerdo). Un hispanohablante habría dicho que los primeros días de estancia en esa ciudad no fueron una experiencia grata, o que no tienen buenos recuerdos de su estancia. Por todo ello, podemos confirmar que se trata de error motivado por transferencia lingüística, pero no motivado por el mandarín, sino por el inglés, la primera lengua extranjera que aprende el alumnado chino. Resulta inevitable comprobar la similitud existente entre la frase mostrada arriba con *I don't have good memories of my first months of stay in that city,* en inglés.

Figura 4
Visualización del décimo-séptimo caso de la averiguación *memoria+* en la pantalla *File View*



Fuente: elaboración propia

Sirva este ejemplo como botón de muestra de las múltiples posibilidades que la lingüística de *corpus* (conjuntamente con las herramientas de análisis de *corpus)* proporciona a la hora de explorar el lenguaje y, en particular, el aprendizaje de una segunda lengua. Se trata de un enfoque distinto al tradicional que nos permite, desde conocer las necesidades de nuestros alumnos, hasta la posibilidad de creación de diccionarios y manuales didácticos especialmente diseñados (en este caso) para el aprendiente *sinohablante*.

3.7. Estudio de asociación de palabras

En un gran número de trabajos se ha demostrado que las *colocaciones* y las unidades multiléxicas (tales como las expresiones idiomáticas) son de difícil comprensión y uso para los discentes (Nesselhauf & Tschichold, 2002). Se trata de un hecho de mayor importancia si el alumno examina textos técnicos de un lenguaje de especialidad en particular, puesto que muchas unidades terminológicas son también multiléxicas (Bowker & Pearson, 2002). La opción *Clusters/N Grams* no sólo ayuda a descubrir las asociaciones de palabras existentes alrededor de una en concreto, sino que también permite, *verbi gratia*, comprobar las preposiciones y los verbos auxiliares que se emplean con la unidad léxica seleccionada.

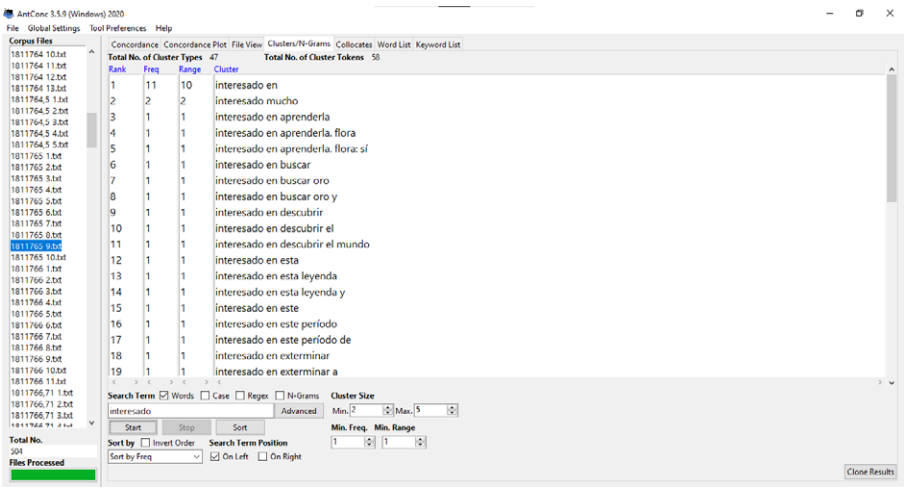
Para una mayor comprensión, en la figura 5 hemos llevado a cabo una búsqueda con el participio del verbo *interesar*. En primer lugar, nuestro interés se enfoca en descubrir qué preposiciones emplean los estudiantes con dicho verbo y, por otro lado, apreciar los verbos auxiliares utilizados. Para ello, nos trasladamos a la pantalla *Clusters/N-Grams* (en la parte superior, sobre fondo blanco), escribimos en el casillero *Search Term* (abajo) la palabra *interesado*, y seleccionamos la casilla *Word* (aunque está preasignada por defecto). Decidimos que la voz elegida se posicione a la izquierda y clicamos en *Search Term Position* la opción *On Left*. Abajo a la derecha, en *Cluster Size*, estimamos idóneo que el programa nos muestre entre un mínimo de dos y un máximo de cinco palabras a la derecha de nuestra voz a indagar.

AntConc nos indica que existen 47 ocasiones en las que aparece el mencionado participio, y que se repite la primera opción en 11 ocasiones (*Freq*, arriba a la izquierda, en color azul). Al haber deseado una indagación por frecuencia de aparición, en *Sort By* (abajo) optamos por mantener *Sort by Freq*. En este caso podemos corroborar que la preposición usada es *en* y el adverbio que aparece es *mucho*, ambos correctos.

Con el propósito de descubrir los verbos auxiliares y adverbios ubicados a la izquierda de *interesado*, optamos porque el vocablo a indagar se posicione a la derecha *(On Right,* en *Search Term Position)*. En esta ocasión aparecen 44 ejemplos, situándose en primer lugar su uso con los adverbios *muy* y *más*, que se repiten en dos ocasiones. De igual modo, corroboramos que el verbo *estar* (además del *haber)* es el que más se emplea, por lo que comprobamos que el estudiantado no suele tener dudas al valerse del susodicho participio (véase figura 6).

Figura 5

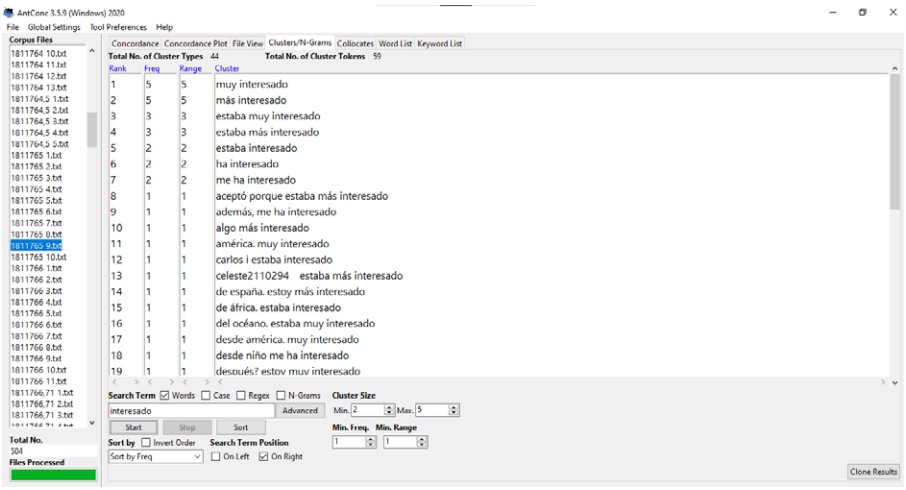
Búsqueda de palabras situadas a la derecha de *interesado*



Fuente: elaboración propia

Figura 6

Búsqueda de palabras situadas a la izquierda de *interesado*



Fuente: elaboración propia

Asimismo, el programa permite mostrar los resultados por orden de frecuencia, alfabéticamente, por promedio, probabilidad, palabra o terminación de ésta, al igual que invertir cada una de las elecciones descritas al pulsar *Invert Order*, ubicado a la derecha de *Sort By* (Anthony, 2004, p. 11). En definitiva, se trata de una novedosa forma de apreciar, con mayor especificidad, ciertos aspectos lingüísticos, sin tener que emplear una búsqueda de concordancia *(Concordance)* y de rehuir de un número excesivo de datos que dificulten o perturben en nuestras pesquisas. En el campo de la terminología, a ese exceso de información no deseada se le denomina *ruido* (Cabré et al., 2001, p. 76), que requiere de tiempo para proceder a su análisis, depuración y eliminación.

3.8. Estudio de colocaciones

Con referencia a los hechos recursivos más proclives a aparecer en una concordancia, el primero de ellos es la *colocación*, que Clear define como "la *coocurrencia* recurrente de las palabras" (Firth, 1957, p. 277). Las *colocaciones* se caracterizan por términos de frecuencia y por su posición, así como por su variación compositiva o idiomática. Dicho de otro modo: una palabra no sólo se asocia con unidades léxicas con clara significación, sino también con marcadores gramaticales, al igual que con diferentes categorías gramaticales (McEnery & Hardie, 2012, p. 130).

Se trata de una extracción similar a la de *Cluster/N Grams*, pero con una mayor precisión sobre las voces que acompañan a la analizada que, en este caso, es *seguridad* (véase figura 7). Nuestra intención es examinar qué verbos, preposiciones, adverbios, etc., se unen con el mencionado sustantivo. Para ello, debemos clicar sobre la pestaña *Collocates* (véase arriba de la pantalla). En esta ocasión aparecen seis columnas. En la primera *(Rank)* se nos indicará la posición que cada *colocación* posee, la segunda *(Freq)* nos indica el número total de ocasiones que aparece en el *corpus,* la tercera, o *Freq(L)*, nos destaca si dicha voz aparece a la izquierda, y la cuarta, o *Freq(R),* a la derecha. En la quinta columna se visualizará el porcentaje estadístico otorgado *(Stat)* y, por último, la *colocación*. Cabe destacar que (como en todas las aplicaciones de AntConc) cada una de estas filas puede ensancharse o reducirse, al acercar y deslizar el ratón en uno de sus bordes.

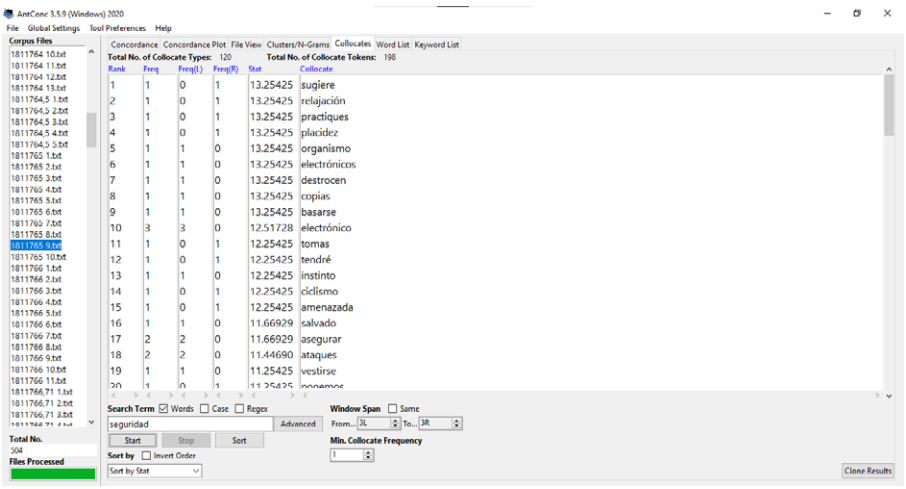
Hemos decidido delimitar nuestra búsqueda a las tres palabras que aparecen tanto a la izquierda (3L) como a la derecha (3R) de la palabra clave, accesible abajo a la derecha, en *Window Span*, *From…To…* Bajo ésta, y al desear

indagar todas las palabras posibles, mantenemos que la frecuencia mínima de la colocación sea de 1 *(Min. Collocate Frequency)*. Como se destaca en la parte superior, a la izquierda, la pesquisa nos arroja una cifra total de 120 palabras diferentes *(Total No. of Collocate Types)*, las cuales aparecen 198 veces alrededor de la palabra clave *(Total No. of Collocate Tokens)*. Como es habitual, si pinchamos sobre cualquiera de las *colocaciones*, el *software* nos trasladará a la ventana *Concordance*, para visualizarla en una línea. Si deseamos analizarla más en profundidad, basta de nuevo clicar sobre la voz para trasladarnos a la *File View* y visionar la *colocación* en el archivo textual donde se encuentra. Por otro lado, aunque hemos seleccionado una confección estadística, es posible ordenar los resultados por frecuencia, dando (si así se desea) una mayor relevancia a los ubicados a la derecha o a la izquierda, por orden alfabético o terminación de palabra, todo ello disponible en la parte inferior izquierda, en *Sort by* (abajo a la izquierda).

En nuestra prospección aparecen en las primeras posiciones los verbos sugerir, practicar, destrozar, basarse, tomar, tener y amenazar. No cabe duda que algunos de estos verbos no suelen tener mucha relación con la seguridad, por lo que un examen más detallado nos indicará si los discentes han empleado correctamente dichos verbos, apreciar los errores (si los hubiere) y dictaminar el motivo de su empleo.

Cada una de estas tres columnas puede ampliarse o reducirse, pudiéndose guardar los resultados de búsqueda alineados, como aparecen en la pantalla en un portapapeles, o trasladarlos a un archivo de texto convencional. AntConc posee esta peculiaridad y, simplemente, clicando sobre la pestaña ubicada en el extremo inferior derecho *(Clone Results)*, nos permitirá guardar los resultados, pudiéndolos contrastar con otros que ejecutemos *a posteriori*.

Figura 7
Examen de las *colocaciones* de *seguridad*



Fuente: elaboración propia

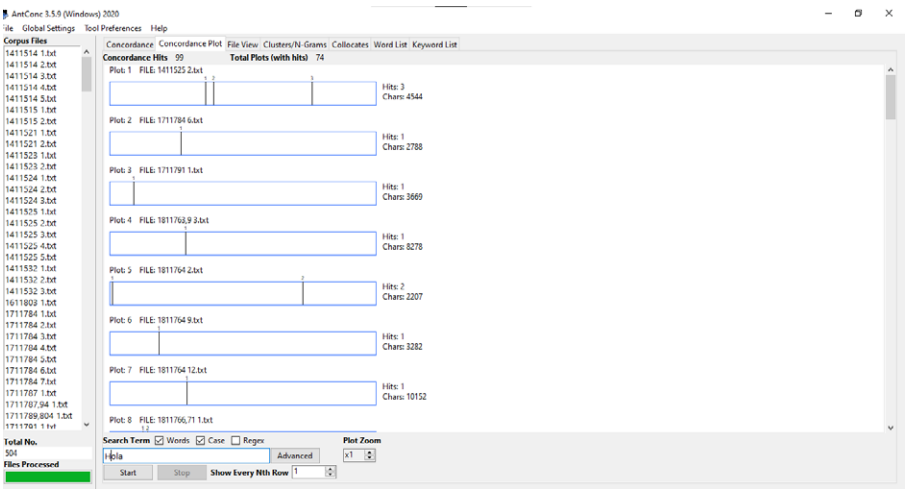
3.9. Ubicación de concordancias

El propósito de este método (denominado en inglés *Concordance Plot)* es mostrar dónde se ubica y cómo se emplea un término/palabra que deseamos analizar dentro de un *corpus* textual. Se trata de una herramienta muy útil tanto en el campo de la terminología como en el científico para, entre otras cosas, identificar la definición de siglas y acrónimos. En términos generales, los términos son detallados al principio del artículo, siendo empleadas sus siglas en el cuerpo del texto. En esta ocasión visualizamos en cada archivo textual donde aparece el vocablo a examinar en un cuadro rectangular, a modo de código de barras, mostrándonos (por medio de una línea) cuándo y dónde aparece la voz a examinar en cada uno de los archivos. En esta ocasión podemos visualizar el número de veces que aparecen y su distribución a lo largo de cada texto (Anthony, 2004, p. 9).

Con un propósito meramente didáctico, hemos decidido averiguar cuántas veces aparece en nuestro *corpus* la interjección *hola* con mayúscula inicial (véase figura 8). En ese caso, escribimos la palabra y seleccionamos las pestañas *Word* y *Case*, a la derecha de *Search Term* (ver abajo). De ese modo, el programa únicamente nos mostrará la interjección escrita con mayúscula

inicial. El motivo es muy simple: en nuestro *corpus* no existen casos de *hola* escrito con minúscula inicial.

Figura 8
Visualización de *Concordance Plot*



Fuente: elaboración propia

Hola se encuentra escrito en 99 ocasiones (*Concordance Hits*, ver arriba), distribuidos en 74 archivos *(Total Plots -with hits-)*. En la pantalla sólo es posible visualizar los primeros ocho textos en donde aparece nuestra palabra clave, por lo que es necesario usar la pestaña que aparece a la derecha de la imagen para descender y analizar el resto. Al posicionarse sobre cualquiera de las barras que aparecen en cada uno de los rectángulos (cada archivo textual donde aparece palabra objeto de análisis), AntConc nos trasladará a dicho archivo y podremos visualizarlo automáticamente en la pantalla *File View*, donde veremos la palabra en contexto, como se puede ver en la figura 4. Se trata, en definitiva, de diferentes maneras de estudiar el lenguaje escrito por medio de una metodología que acerca el estudio del lenguaje a la epistemología científica.

En suma, el texto posee una función que no sólo queda patente en un *contexto* verbal, sino que se expande hacia un contexto específico situacional y cultural. Ante todo, a un archivo textual se le interpreta por su funcionalidad. Los parámetros empleados para el análisis de *corpus* son, ante todo, formales,

y el tipo de información extraído de un texto se considera *significativo*. Los datos recopilados de un *corpus* pueden catalogarse de ese modo, puesto que pueden generalizarse al lenguaje en general, aunque sin poseer una conexión directa con un caso o ejemplo en particular (Tognini-Bonelli, 2001, p. 3).

Por otro lado, un buen número de académicos considera la lingüística de *corpus* como una mera metodología y no como una disciplina. McEnery, Xiao y Tono aducen que se trata de un sistema de métodos y principios de cómo aplicar los *corpus* en estudios lingüísticos y que, de hecho, poseen postulado teórico. No obstante, también aducen que puede considerarse como una metodología con un amplio abanico de aplicaciones en diferentes áreas y teorías lingüísticas (2006, pp. 7-8). Independientemente de la opinión que se tenga al respecto, resulta evidente lo mucho que la lingüística de *corpus* aporta a la mejora en el aprendizaje y enseñanza de una segunda lengua.

4. Discusión: el uso de la lingüística de *corpus* en el aula

El objetivo primordial de este trabajo es enseñar a emplear la lingüística de *corpus* en el aula. No se trata de instruir al alumnado en el uso de las herramientas de gestión de *corpus* ni las premisas teóricas que dicha metodología entraña, sino de ayudarles a ser autosuficientes para analizar y contrastar sus errores con el uso que hacen los nativos de la lengua española. El empleo de la lingüística de *corpus* en el aula no es un hecho que entrañe excesiva dificultad (Bennett, 2010) y, con el propósito de facilitar su comprensión, procederemos a exponer un ejemplo práctico sobre su uso.

Una vez analizado un error representativo entre los trabajos redactados por los aprendientes, utilizamos AntConc para exponerlo en clase por medio de la pantalla de *Concordance*, donde se detallan las concordancias o casos resultantes de una búsqueda cualquiera. *Verbi gratia*, hemos seleccionado una voz empleada erróneamente por los aprendientes chinos, que es *pensamiento*. En mandarín, *sīxiǎng* (思想) se traduce como *ideología, idea* o *mente*. No obstante, algunos aprendientes tienden a confundir dichos conceptos al concebirlo todo *(grosso modo)* como un *acto de pensar*; de ahí que empleen *pensamiento* cuando se refieren a ideas, puntos de vista, formas de pensar o incluso a la experiencia adquirida por las personas.

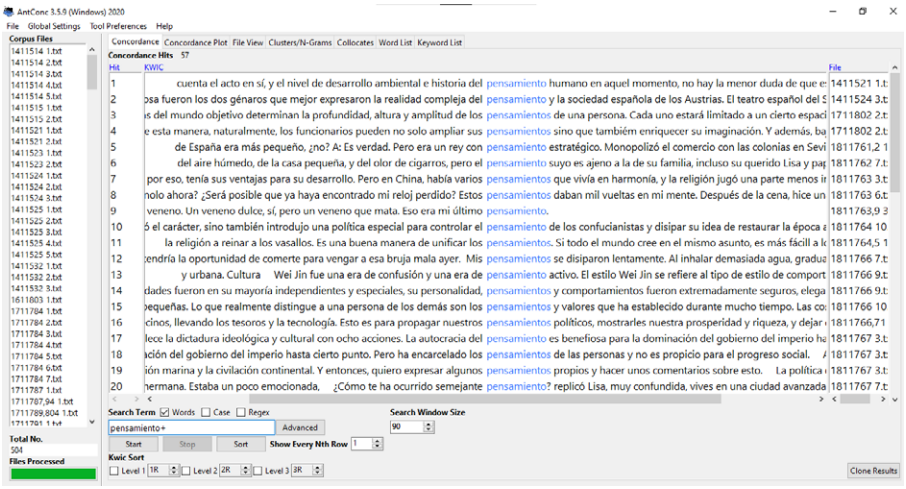
En la figura 9 observamos que han aparecido 57 resultados en la búsqueda de *pensamiento+* (a saber, su forma en singular más una letra adicional, para incluir también su forma en plural) en nuestro *corpus* de aprendientes. Entre

ellos, podemos ver que su uso, en algunas ocasiones, no es del todo acertado, al emplearse en frases como *…altura y amplitud de los pensamientos de una persona* (*Hit 3*), *…los funcionarios pueden no sólo ampliar sus pensamientos… (Hit 4)*, *… Pero era un rey con pensamiento estratégico (Hit 5)* o *…quiero expresar algunos pensamientos propios… (Hit 19)*.

Para que los aprendientes comprueben el uso de la mencionada voz por parte de los hispanohablantes, usamos un *corpus* de referencia español. A la hora de analizar los datos extraídos de un *corpus* es necesario compararlos con un repositorio textual de mayores dimensiones, llevado a cabo por hablantes nativos y con una amplia variedad de géneros. A ese *corpus* de mayores dimensiones se le denomina *corpus de referencia*, que no simboliza ninguna variedad lingüística, de registro o textual, sino que se ha compilado con el propósito de representar una lengua en particular al albergar la mayor cantidad de géneros textuales posibles (Baker et al., 2006, pp. 137-138). Atendiendo a esas razones, hemos seleccionado el Corpus de Referencia del Español Actual o CREA[4] (Real Academia Española, 2008).

4. El *corpus* CREA es uno de los diferentes repositorios textuales y transcritos de la Real Academia Española. Contiene textos procedentes de todos los países de habla hispana y de una variedad temática que le proporciona un incalculable valor en el estudio e investigación de nuestra lengua. Se encuentra disponible en el siguiente enlace: http://corpus.rae.es/creanet.html

Figura 9
Resultados de la búsqueda de *pensamiento+*



Fuente: elaboración propia

El uso de un *corpus* de referencia en el aula evita el empleo de la introspección y proporciona ejemplos del uso real de una lengua determinada. En este caso, y debido al elevado número de ejemplos de *pensamiento*, nos hemos limitado a mostrar en la figura 10 algunos de los resultados obtenidos en el campo de lingüística y lenguaje. Al analizar los datos obtenidos en CREA, el alumnado apreciará el uso correcto de dicha voz en particular y cómo, en ocasiones, es más adecuado traducirla al español como *idea, concepto, mente, conocimiento, experiencia* o *punto de vista*. Para ello, procederemos seguidamente a llevar a cabo una búsqueda de las mencionadas palabras y expresiones para que los aprendientes puedan apreciar cómo se utilizan, al igual que las diferentes connotaciones existentes entre dichas equivalencias. Al descubrir y habituarse al empleo didáctico del Corpus de Referencia del Español Actual, el aprendiente será capaz de consultar cualquier palabra o expresión en dicho repositorio para así mejorar su lenguaje y aclarar sus dudas, mejorando el nivel que posee de español.

Figura 10
Visualización del *corpus* de referencia del español actual

Fuente: elaboración propia

5. Conclusiones

En este trabajo hemos descrito someramente qué es la lingüística de *corpus*, los *corpus* (en particular, el de aprendientes), su compilación y las herramientas a emplear para su análisis, en concreto el programa AntConc, versión 3.5.9. (Windows) 2020. Además de detallar su instalación y empleo por medio de ejemplos; también hemos pormenorizado los motivos que nos han llevado a compilar el CEEC y los objetivos que deseamos alcanzar. Seguidamente, hemos detallado un ejemplo práctico de cómo implementar la inserción de la lingüística de *corpus* en el aula mediante la pormenorización de errores con AntConc y su análisis con el Corpus de Referencia del Español Actual. Esta investigación se ha enfocado en los factores específicos que afectan a los discentes chinos al aprender la lengua española, en particular en los errores que éstos cometen motivados por las interferencias lingüísticas existentes en su lengua materna, el español y el inglés (Armas, 2018, p. 160). Se trata de un objetivo limitado, que deseamos continuar y ampliar, de la misma manera en que se amplía nuestro *corpus*.

Finalmente, destacar que el propósito que ha motivado la redacción de este artículo es dar a conocer esta metodología (o disciplina), invitar a nuestros colegas a utilizarla y animarlos a que practiquen en el uso de las herramientas

de *corpus;* de ese modo, podrán apreciar las carencias y dificultades a las que se enfrentan sus aprendientes, mediante un enfoque científico. Al incorporar la lingüística de *corpus* en el aula como material didáctico, el aprendiente será capaz de identificar los errores que él mismo comete, mediante un proceso de razonamiento inductivo (Baker et al., 2006, pp. 102-103).

Referencias

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *IWLeL 2004: An interactive Workshop on Language e-Learning*, 7-3. https://waseda.repo.nii.ac.jp/records/28823

Armas, A. R. (2018). Mejora de la competencia ortográfica en el plano escrito a través de un taller de escritura dirigido a universitarios sinohablantes. *Monográficos Sinoele*, (17), 157-70.

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics*. Edinburgh University Press. https://doi.org/10.1515/9780748626908

Bennett, G. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. University of Michigan Press. https://doi.org/10.3998/mpub.371534

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language. A Practical Guide to Using Corpora*. Routledge. https://doi.org/10.4324/9780203469255

Cabré, M. T., Estopà, R., & Vivaldi, J. (2001). Automatic term detection. A review of current systems. En D. Bourigault, C. Jacquemin & M.C. L´Homme (Eds.), *Recent Advances in Computational Technology* (pp. 53-87). John Benjamins.

Castillejos, W. (2009). Error analysis in a learner corpus: What are the learners' strategies. En P. Cantos & A. Sánchez (Eds.), *A Survey of Corpus-Based Research* (pp. 675-690). Murcia: Asociación Española de Lingüística del Corpus.

Cerezo, E. (2014). Análisis curricular de la formación lingüística y sociocultural en las titulaciones de Traducción e Interpretación en España. *Prosopopeya: revista de crítica contemporánea*, (9), 283-315.

Cerezo, E. (2019). Lenguas extranjeras con fines traductológicos: en busca de una identidad propia. *Quaderns. Revista de Traducció*, (26), 239-254.

Ellis, R. (2003). *The Study of Second Language Acquisition*. Oxford University Press.

Els, T. V. (1984). *Applied Linguistics and the Learning and Teaching of Foreign Languages* (R. R. van Oirsouw, Trad.). Hodder Arnold.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. Oxford University Press.

Gamon, M., Chodorow, M., Leacock, C., & Tetreault, J. (2013). Using learner corpora for automatic error detection and correction. En A. Díaz-Negrillo, N. Ballier & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 127-150). John Benjamins. https://doi.org/10.1075/scl.59.09gam

Hurtado Albir, A. (1999). Objetivos de aprendizaje y metodología en la formación de traductores e intérpretes. En A. Hurtado (Coord.), *Enseñar a traducir: metodología en la formación de traductores e intérpretes* (pp. 8-58). Edelsa.

James, C. (1992). Awareness, consciousness and language contrast. In C. Mair & M. Markus (Eds.), *New departures in Contrastive Linguistics* (pp. 183–198). University of Innsbruck Press.

James, C. (1998). *Errors in Language Learning and Use: Exploring Error Analysis*. Routledge. https://doi.org/10.4324/9781315842912

Leech, G. (1992). Corpora and Theories of Linguistic Performance. En J. Svartvik (Ed.), *Directions in Corpus Linguistics: Proceedings of the Nobel symposium 82* (pp. 105-122). Mouton de Gruyter.

Loffler, S. (2011). Las interferencias lingüísticas y la enseñanza del español como lengua extranjera. *ARJÉ Revista de Posgrado FACE-UV, 5*(8), 11-25. http://www.arje.bc.uc.edu.ve/arj08/art01.pdf

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. Taylor & Francis.

Millán, R. (2006). Interferencias lingüísticas en el aprendizaje de una segunda lengua. En A. Álvarez, L. Barrientos, M. Braña, V. Coto, M. Cuevas, C. de la Hoz, I. Iglesias, P. Martínez, M. Prieto & A. Turza (Eds.), *La competencia pragmática y la enseñanza del español como lengua extranjera: Actas XVI Internacional ASELE* (pp. 481-485). Ediciones de la Universidad de Oviedo.

Nesselhauf, N., & Tschichold, C. (2002). Collocations in CALL: An investigation of vocabulary-building software for EFL. *Computer Assisted Language Learning*, *15*(3), 251-279. https://doi.org/10.1076/call.15.3.251.8190

Real Academia Española. (2018). *Corpus de Referencia del Español Actual*. RAE. https://www.rae.es/banco-de-datos/crea

Selinker, L. (1974). Interlanguage. En J. C. Richards (Ed.), *Error Analysis: Perspectives on Second Language Acquisition* (pp. 31-54). Longman.

Selinker, L., & Rutherford, W. E. (2013). *Rediscovering Interlanguage*. Routledge.

Sinclair, J. (2005). Corpus and Text - Basic Principles. En M.Wynne (Ed.), D*eveloping Linguistic Corpora: A Guide to Good Practice*, (pp. 1-16). AHDS.

Swanson, B., & Charniak, E. (2014). Data Driven Language Transfer Hypotheses. En *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, (Vol. 2, pp. 169-173). Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-4033

Tetreault, J., Blanchard, D., & Cahill, A. (2013). A Report on the First Native Language Identification Shared Task. En J. Tetreault, J. Burstein & C. Leacock (Eds.), *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 48-57). Association for Computational Linguistics. https://aclanthology.org/W13-1706.pdf

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins. https://doi.org/10.1075/scl.6

Wang, M., Malmasi, S. & Huang, M. (2015). The Jinan Chinese Learner Corpus. En J. Tetreault, J. Burstein & C. Leacock (Eds.), *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 118-123). Association for Computational Linguistics. https://aclanthology.org/W15-0614.pdf

Weisser, M. (2016). *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. Wiley.

Yip, V. (1995). *Interlanguage and Learnability: From Chinese to English*. John Benjamins. https://doi.org/10.1075/lald.11