Diseño de tareas y materiales para recopilar un corpus de aprendices mexicanos de lengua inglesa

Design of tasks and materials to collect a corpus of Mexican learners of English

Ana Abigahil Flores Hernández
Universidad Autónoma del Estado de México, Facultad de Lenguas
aafloresh@uaemex.mx

Pauline Moore
Universidad Autónoma del Estado de México, Facultad de Lenguas
pmooreh@uaemex.mx

Recepción: 26 de abril del 2023
Aceptación: 22 de agosto del 2023
doi: 10.22201/enallt.01852647p.2024.78.1060

Resumen

Este trabajo describe el proceso de diseño de una entrevista oral para recopilar datos de producción lingüística de aprendices de inglés como segunda lengua en universidades públicas mexicanas. El objetivo de la aplicación de este instrumento es crear un corpus de aprendices variado y representativo que sirva como base para investigaciones en adquisición de segunda lengua, así como para el diseño de materiales y estrategias de enseñanza acordes a las necesidades de los estudiantes mexicanos. Para ello se realizó una rigurosa selección de tareas y diseño de materiales con base en dos conceptos clave para el diseño de un corpus de aprendices: la autenticidad y la representatividad. El resultado es una entrevista de cuatro secciones alineadas a las actividades productivas orales monológicas de acuerdo con el Marco Común Europeo de Referencia para las lenguas (monólogo descriptivo, informativo y argumentativo). Las tareas además representan tres de los géneros textuales de acuerdo con el modelo de análisis multidimensional para textos orales: textos informativos, narrativos y de posicionamiento.

Palabras clave: lingüística de corpus; corpus de aprendices; adquisición de L2; enseñanza de L2; lenguas extranjeras

Abstract

This paper describes the design process of an oral interview to be used in the collection of language-production data of university learners of English as a second language. The aim of applying this instrument is to build a relatively large and representative corpus of learners to be used in research about second-language acquisition in Mexico and for the creation of English language teaching strategies and programs consistent with the particular characteristics of Mexican students. To this end, we carried out rigorous task selection and materials design during this instrument design process to ensure two key aspects in the construction of the learner corpus: authenticity and representativeness. This process resulted in a four-part oral interview aligned to the monologic spoken productive activities according to the Common European Framework of Reference for Languages, namely descriptive, informative, and argumentative. These tasks also represent three text types according to the multidimensional analysis for oral texts, information-oriented, stance-oriented, and narrative.

Keywords: corpus linguistics; learner corpora; second language acquisition; language teaching; foreign languages

1. Introducción*

La lingüística de corpus se ha convertido en una fuente de evidencia para complementar datos obtenidos con técnicas de elicitación e introspección. Esta metodología permite visualizar lo que ocurre en el lenguaje y analizarlo desde el paradigma cuantitativo para entender los mecanismos de adquisición de una segunda lengua y mejorar, sobre esta base, la enseñanza de segundas lenguas (Granger, 2002; 2008). Hasta ahora, el diseño de programas y materiales de enseñanza de la lengua inglesa ha partido de un modelo que considera el análisis de necesidades desde dos aspectos: las características de los aprendices (estilos, necesidades, aptitudes, etc.) y las descripciones de la lengua (estructuras, funciones comunicativas, etc.). Ahora bien, los corpus de aprendices permiten agregar un tercer aspecto: la producción de quienes adquieren la segunda lengua. Esta producción brinda evidencia empírica del proceso de adquisición de la lengua y de diversos factores que influyen en él, al mismo tiempo que representa el producto de este proceso y de la influencia de dichos factores (Granger, 2002).

De acuerdo con el listado *Learner corpora around the world* (Centre for English Corpus Linguistics, 2020), que almacena las referencias y sitios web de los corpus de aprendices en el mundo, son pocas las colecciones dedicadas a recopilar la producción de aprendices de inglés que tienen como lengua materna el español (seis corpus en total) y ninguna de ellas se ha dado a la tarea de recopilar producción de aprendices mexicanos. Por otro lado, existen colecciones amplias, recopiladas en su mayor parte por grandes editoriales enfocadas en la enseñanza del inglés, que tienen como finalidad proveer material para obras de referencia y libros de texto para la enseñanza del inglés como lengua extranjera, las cuales contienen un porcentaje mínimo de representación de la variante

* Las autoras desean expresar su agradecimiento al Conahcyt por la beca otorgada a Ana Abigahil Flores Hernández (Convocatoria de Estancias Posdoctorales 2020, 2021 y 2022).

mexicana. Sin embargo, algunas de estas colecciones son de acceso restringido por tratarse de un recurso propiedad del sello editorial, otras dificultan la consulta directa de la producción de los aprendices mexicanos y otras solo incluyen producción escrita.

El corpus de aprendices mexicanos (Mexican Learner Corpus o MexLeC) constituye un proyecto que recolecta de manera longitudinal la producción oral de aprendices mexicanos de inglés como segunda lengua, utilizando una entrevista guiada conformada por cuatro diferentes tareas comunicativas. Los datos obtenidos con este instrumento pretenden conformar un robusto corpus de aprendices que permita, además de las metas inmediatas de monitorear el progreso de la producción oral, desarrollar estrategias y materiales específicos para los participantes y servir como base para un amplio rango de investigaciones sobre adquisición y enseñanza de segundas lenguas.

Para lograr estos objetivos, se ha diseñado un instrumento confiable y válido que permite recopilar suficientes ejemplares de producción lingüística, seleccionando con rigor cada una de las tareas y materiales incluidos para guiar tanto a los participantes como a los aplicadores y lograr una estandarización en la aplicación del instrumento, así como homogeneidad en los datos recabados con base en los conceptos de *autenticidad* y *representatividad*. La descripción amplia y detallada de este proceso de selección y diseño de tareas es el objetivo de este trabajo.

2. Corpus de aprendices: definición y tipología

Un área de auge en la lingüística actual es la investigación del uso de la lengua mediante los corpus electrónicos, que complementa la investigación basada en la intuición y juicios de los hablantes nativos (Granger, 2021). La lingüística de corpus es una metodología que implica el uso de corpus en investigaciones cualitativas o cuantitativas sobre la lengua por medio de su procesamiento y haciendo uso de *software* especializado, previa implementación de una normalización o etiquetado que posibilite el manejo de gran-

des volúmenes de datos y la replicabilidad de las investigaciones (McEnery & Gabrielatos, 2006; Lehmberg & Wörner, 2008; McEnery & Hardie, 2011).

Un corpus es una colección grande de textos auténticos y representativos de una cierta variante de lengua oral o escrita, que se encuentra almacenada en formato electrónico (Baker, Hardie & McEnery, 2006; Lüdeling & Kytö, 2008). Esta misma definición puede aplicarse para un corpus de aprendices, con la diferencia de que los textos recopilados son producidos por aprendientes de una segunda lengua (Granger, 2008). Así, un corpus de aprendices puede definirse como una colección sistemática de lengua continua, contextualizada y auténtica (escrita o hablada) que se almacena en formato electrónico y constituye una clase especial de datos empíricos que pueden usarse en diferentes disciplinas relacionadas con el aprendizaje de segundas lenguas, el diseño de materiales pedagógicos y el desarrollo de herramientas de entrenamiento y procesamiento de lenguaje natural (Granger, 2008; Callies & Paquot, 2015; Meunier, 2021).

Los corpus de aprendices pueden clasificarse según seis dimensiones: el modo, el género, la lengua meta, la lengua materna, el tiempo de recolección, su alcance y su propósito (Sinclair, 2005; Gilquin, 2015). El modo se refiere al contexto de recopilación y almacenaje de los datos, de acuerdo con este criterio un corpus puede ser oral o escrito, presencial o virtual y monológico o dialógico (Sinclair, 2005). Este criterio incluye el tipo de almacenaje, que se refiere a la clase de archivo en el que se guardan los datos. En específico, los corpus pueden ser textuales orales o callados, lo que implica, en el primer caso, el uso de archivos de texto, archivos de audio y video (con sus transcripciones), o bien, en el segundo caso, solo archivos de texto o transcripciones que representan la producción oral.

El género de un corpus hace referencia al contexto de uso de lengua o el propósito de las producciones que conforman la colección. Así, hay corpus de lengua con propósitos específicos, por ejemplo, de negocios o académicos, o cuyos textos representan

géneros discursivos, por ejemplo, cartas formales o informales, ensayos, narrativas, entre otros. Por último, a los corpus que reúnen varios géneros se les denomina corpus de lengua generales.

La lengua meta de un corpus es la utilizada por los participantes durante el proceso de recolección. Los corpus de aprendices se definen según dos lenguas. Por un lado, se considera la L2 que se está aprendiendo, en la que los participantes realizan su producción. Por otro lado, la lengua materna de los aprendices también juega un papel fundamental, debido a que influye en las expresiones y estructuras lingüísticas producidas.

Por el tiempo de recolección, un corpus puede ser sincrónico o diacrónico, es decir, puede ser recolectado en un solo punto en el tiempo (corpus transversal), o puede obtenerse mediante diferentes recolecciones con el fin de representar la evolución de la lengua de los aprendices a través del tiempo (corpus longitudinal). Según su alcance, un corpus puede ser local o global. Es global si contiene datos de muchos sujetos de diversas regiones o países del mundo y es local si los textos son recolectados para representar una población específica, por ejemplo, estudiantes de una universidad o pertenecientes a un curso o programa educativo específico. Además, los corpus pueden clasificarse en torno a su origen y propósito, es decir, el uso que se da a los datos recolectados. En este sentido pueden clasificarse en dos categorías, la primera, con fines de investigación y de acceso libre (o casi libre) para la comunidad científica y la segunda, de uso reservado, cuya finalidad es el diseño de materiales de enseñanza comercializables.

El corpus MexLeC, objeto de este trabajo, es una colección que puede considerarse en cuanto al modo como un corpus oral, virtual, monológico y callado, ya que representa la producción oral de aprendices que son entrevistados en modalidad virtual mediante tareas monológicas que después se transcriben y almacenan en archivos de texto (y no en archivos de audio). La lengua meta del corpus es inglés como segunda lengua y se trata de un corpus de aprendices general, pues no se circunscribe a algún género o variante lingüística específica. Respecto al tiempo de recolección,

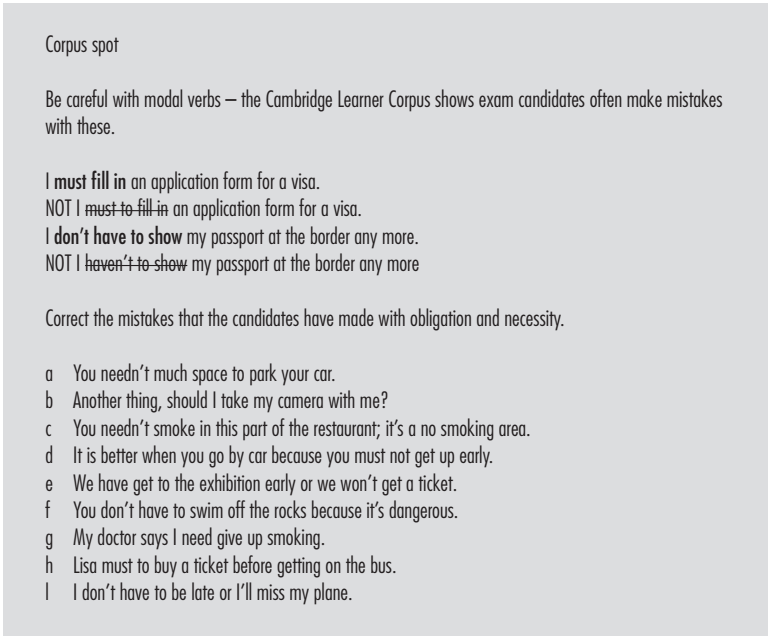
constituye un corpus diacrónico, pues plantea un seguimiento longitudinal del desarrollo de la lengua de cada participante durante cuatro o cinco años. En cuanto a su cobertura, se considera de alcance nacional, ya que recopila producciones de estudiantes de diversas universidades en México que se especializan en lenguas modernas. Por último, MexLeC es un corpus de uso libre con fines de investigación o académicos.

Los corpus de aprendices más notables por su tamaño y sus innumerables aplicaciones en la investigación sobre adquisición y enseñanza de segunda lengua son cuatro, dos comerciales y dos académicos. Por un lado, los corpus comerciales incluyen el Cambridge Learner Corpus, que contiene más de 1800 millones de palabras producidas durante las certificaciones de Cambridge ESOL, y el Longman Learners' Corpus, integrado por 10 millones de palabras representativas de aprendices de 20 lenguas maternas distintas (Meyer, 2002). Ambos fueron creados con el objetivo de sustentar materiales comercializables de enseñanza y aprendizaje de inglés, y han sido utilizados con frecuencia en la creación de libros de preparación para certificaciones y diccionarios bilingües para hablantes no nativos.

Por otro lado, entre los corpus académicos se encuentran el International Corpus of Learner English (Granger, Dupont, Meunier, Naets & Paquot, 2020) y el Trinity Lancaster Corpus (Gablasova, Brezina & McEnery, 2019), ambos con un promedio de 4.5 millones de palabras. Estas colecciones tienen como objetivo común conformar una base para la investigación sobre adquisición y enseñanza del inglés como segunda lengua. Estos corpus son de acceso libre y se han empleado en estudios sobre adquisición de L2, por ejemplo, Gilquin (2019), que realiza una exploración sobre el uso de verbos ligeros como *take* o *make* en el Trinity Lancaster Corpus.

3. Utilidad de un corpus de aprendices

En el campo de la enseñanza de lenguas extranjeras, los datos de un corpus de aprendices representan una aportación valiosa, sobre



Corpus spot

Be careful with modal verbs – the Cambridge Learner Corpus shows exam candidates often make mistakes with these.

I **must fill in** an application form for a visa.
NOT I ~~must to fill in~~ an application form for a visa.
I **don't have to show** my passport at the border any more.
NOT I ~~haven't to show~~ my passport at the border any more

Correct the mistakes that the candidates have made with obligation and necessity.

a You needn't much space to park your car.
b Another thing, should I take my camera with me?
c You needn't smoke in this part of the restaurant; it's a no smoking area.
d It is better when you go by car because you must not get up early.
e We have get to the exhibition early or we won't get a ticket.
f You don't have to swim off the rocks because it's dangerous.
g My doctor says I need give up smoking.
h Lisa must to buy a ticket before getting on the bus.
l I don't have to be late or I'll miss my plane.

FIGURA 1. Ejercicio basado en el análisis de errores según un corpus de aprendices (Chapelle & Sharp, 2014)

todo para la creación de materiales pedagógicos y el diseño de programas e instrumentos de evaluación (Granger, 2008). El uso de un corpus auténtico que muestra los errores comunes y los conocimientos adquiridos por los aprendientes permite que los diseñadores puedan enfocarse en necesidades reales de aprendizaje. Entre los resultados de estas aportaciones se encuentran los diccionarios para aprendices, como el *Longman Wordwise Dictionary* o el *Longman Dictionary of Contemporary English*, ambos basados en el Longman Learners' Corpus. El análisis de los corpus de aprendices permite que el lexicógrafo pueda comentar en las entradas léxicas problemas de uso que presentan los aprendices. Por ejemplo, en la entrada para *clothes* del *Longman Wordwise Dictionary* se encuentra la acotación de uso mostrada en (1).

(1) *Do not use* ***cloth*** *or* ***cloths*** *to mean "the things that people wear". Instead use* ***clothes****, a clothes shop.* | *The guests all wore casual clothes*.
'no utilice *cloth* o *cloths* para referirse a lo que la gente viste. En lugar de ello use *clothes*, *a clothes shop.* | Los invitados usaban ropa casual'

Otras aplicaciones incluyen el diseño de materiales para enseñanza que abordan errores usuales de los aprendices (Granger, 2002). La editorial Cambridge University Press incluye contenidos de práctica en sus libros de preparación para exámenes de certificación derivados de su corpus de aprendices, el Cambridge Learner Corpus. En la Figura 1 se muestra un ejercicio tomado de Capel y Sharp (2014), que presenta los errores más comunes en el uso de los verbos modales *must* y *have to*, según el Cambridge Learner Corpus.

4. Consideraciones en el diseño de un corpus

El diseño de un corpus tiene una relación directa con el objetivo de uso de la colección. En el caso de un corpus general es necesario definir con cuidado el contenido con respecto a las personas que participan y sus características, los materiales, el medio de recolección y el método de transcripción elegido (Hunston, 2008). Meunier (2021) considera que los criterios principales para seleccionar los textos a incluir en un corpus de aprendices son: 1) los aprendices y sus características, su lengua materna y sus niveles de dominio de la lengua meta; 2) los tipos de tareas seleccionadas y su propósito, ya sean orales, escritas, formales, informales, expositivas, narrativas, entre otras, y 3) las condiciones en las que se producen estas tareas: interactivas, en ambientes virtuales o como parte de una evaluación. Estos criterios pretenden orientar las decisiones iniciales del proceso de diseño y obedecen a las características fundamentales que debe cumplir un corpus lingüístico: autenticidad y representatividad (McCarthy & O'Keeffe, 2009), así como practicidad y conveniencia para el investigador (McEnery & Hardie, 2011).

4.1. *Autenticidad*

La *autenticidad* en un corpus se refiere a que las muestras de la lengua contenida hayan sido producidas por los hablantes en contextos naturales (McEnery & Hardie, 2011), es decir, que se trate de una comunicación genuina generada en contextos reales y situaciones cotidianas, no artificiales o forzadas (Sinclair, 1996). En el caso de un corpus de aprendices, la producción incluida no podría ser considerada auténtica en el sentido estricto del concepto. El uso de una L2 como lengua extranjera se reduce al ambiente escolar y a situaciones de aprendizaje en el salón de clases —tareas específicas que deben realizar los aprendices—, de tal suerte que la producción lingüística casi siempre es el resultado de una solicitud de un hablante más competente, sea este el maestro o el entrevistador.

A pesar de lo anterior, la lengua incluida en un corpus de aprendices, si bien no se considera del todo auténtica o natural, puede ser vista como "quasi-natural" o "quasi-auténtica" (Meunier, 2021). De acuerdo con Gilquin (2015) y Granger (2008), la lengua producida por aprendices puede clasificarse según diferentes grados de naturalidad, dependiendo de la tarea en que participan los aprendices y el grado de edición o espontaneidad del discurso producido. En este sentido, un corpus oral, a diferencia de un corpus escrito, podría considerarse como una producción más auténtica o más representativa del interlenguaje de los aprendices, debido a que es más espontánea, transcurre en tiempo real y no da tiempo a los procesos de monitoreo o edición que se observan en la producción escrita.

Asimismo, existen niveles de naturalidad que reflejan el tipo de datos producidos en relación con las tareas utilizadas para estimular dicha producción. Así, las tareas "más naturales" son aquellas que se enfocan en la comunicación de significados, en las cuales el aprendiz elige las palabras que usará para expresarse. En estas el participante no es obligado a producir estructuras o vocabulario específicos, ya que generan cierta clase de comunicación cuyas estructuras y vocabulario son impredecibles (Granger, 2008). Wich-

mann (2008) define el grado de naturalidad del discurso producido a partir de una tarea de acuerdo con el grado de restricciones en las estructuras requeridas en la respuesta. Considerando lo anterior, se puede concluir que un tipo de tarea muy poco natural podría ser la lectura en voz alta o la repetición, mientras que las más naturales parten de la solicitud de opiniones o recuentos de experiencias personales.

En resumen, la autenticidad de un corpus de aprendices dependerá de la selección de tareas que privilegien la comunicación y no la generación de estructuras específicas, y que se realicen dentro de un contexto que no permita la preparación previa, el uso de materiales de apoyo o la edición, con el fin de que se produzca un discurso con un mayor grado de espontaneidad que permita el acceso al repertorio léxico real del aprendiz.

4.2. *Representatividad*

La *representatividad* se define a partir de que el contenido de una colección determinada sea una muestra en menor escala de la variedad de lengua que se pretende documentar. Esto a fin de que su análisis y resultados puedan generalizarse para la totalidad de la variedad lingüística representada. Se puede afirmar que un corpus general y robusto es representativo en la medida en que captura la variabilidad de la lengua (Baker *et al*., 2006; McEnery, Xiao & Tono, 2006). En este caso, la lengua que se desea representar es la producida por aprendices de inglés, la representatividad, entonces, estará determinada por la selección de los bloques de textos incluidos en la colección (corpus), que deberá de haber sido conformada de modo equilibrado.

Según Biber (1993), son dos los criterios para determinar la representatividad: el criterio externo o situacional, definido por el contexto o el tipo de discurso de la lengua producida y que da lugar a una clasificación de géneros textuales (novela, artículo periodístico o noticia); y el criterio interno, determinado por los rasgos lingüísticos específicos de la lengua producida (verbos en infiniti-

vo, tercera persona, modales), que resulta en una clasificación de tipos textuales (narrativos, informativos, o persuasivos, entre otros).

En un corpus de aprendices los géneros textuales corresponden a la variedad de tareas, ya que no se cuenta con una gran cantidad de producciones que concuerden con los géneros textuales de un corpus de hablantes nativos, como textos periodísticos, poemas o entradas de blog. De la misma manera, la variabilidad de la lengua es determinada por las tareas que realizan los aprendices en el salón de clases, así como los materiales diseñados para su aplicación, que influyen de manera decisiva sobre el tipo de discurso generado.

McCarthy y O'Keeffe (2009) señalan que las tareas orales pueden clasificarse en dos tipos: dialógicas o monológicas, dependiendo si hay una interacción con uno o varios interlocutores o si el participante genera un discurso sin intervención de un interlocutor. Wichmann (2008) agrega entre estos dos tipos un nivel intermedio, el diálogo asimétrico o monólogo dirigido, en el cual existe un interlocutor que solo funciona como guía; en general se espera que la mayor parte del discurso sea producido por el participante.

De acuerdo con el listado de corpus de aprendientes en el mundo (Centre for English Corpus Linguistics, 2020), existen alrededor de 40 corpus de aprendientes de lengua inglesa como L2 que contienen textos orales. Considerando las tres categorías de tareas orales antes mencionadas, las de tipo monológico incluyen la repetición de palabras o frases, narraciones, descripciones, presentaciones, lectura en voz alta y preguntas de opinión que no contemplen una interacción con el entrevistador o un par. Las tareas dialógicas constituyen conversaciones, tareas de discusión, juegos de roles o interacciones áulicas que se desarrollan entre participantes. Por último, las tareas que representan un diálogo asimétrico incluyen ciertas preguntas de opinión, conversaciones y entrevistas en las que el participante interactúa con un examinador.

Se puede decir, entonces, que el criterio externo en un corpus de aprendices está determinado por el tipo de tareas que realizan

los aprendices, pues dependiendo de ello se producen distintos géneros textuales, como la narrativa, la descripción, la argumentación, así como diversas formas de discurso: espontáneo, ensayado o repetitivo. Todos estos tipos están centrados en el intercambio comunicativo, la entrega de información o bien, la simple producción oral de figuras acústicas. La representatividad entonces estará delimitada por el tipo de tarea que se seleccione y las condiciones en las que se aplique.

Con respecto al criterio interno de la representatividad, en un corpus oral existen cuatro tipos textuales definidos con base en los rasgos lingüísticos característicos del discurso producido. De acuerdo con el análisis multidimensional (MDA) para los textos orales (Biber, 2004), estos tipos son: discurso orientado a la interacción, discurso de posicionamiento, discurso orientado a la información y discurso narrativo.

El primer tipo textual está orientado a la interacción, y se caracteriza por intercambios con turnos cortos cuya temática refiere aspectos cotidianos sociales o relacionales. En términos semánticos, en este tipo de textos se espera que la mayoría de los verbos sean de actividad, como *walk*, *work*, *create*, y que estén conjugados en el tiempo presente y en primera y segunda persona por enfocarse en los interlocutores.

El segundo tipo es el texto con orientación al posicionamiento, cuya interacción se enfoca en la expresión o intercambio de opiniones. En este caso, se espera que los verbos utilizados pertenezcan a la categoría semántica de verbos mentales, como *think* y *believe*, o factivos, como *be* y *seem*, además, se espera toda la gama de expresiones de modalización, tanto el uso de verbos modales como de procesos de atenuación e intensificación mediante expresiones evasivas, como *probably*, *I am not sure that..., it is very true that...* Debido a que la modalización es prosódica (Hood, 2010), también pueden observarse mezclas de expresiones modalizadas con verbos factivos, esto para generar significados más o menos atenuados como en (2).

(2) *Actually*, *I don't think that* *could* *be true*
FACTIVO MENTAL MODAL FACTIVO
'De hecho, no creo que eso pueda ser verdad'

Dentro de la clasificación de textos orientados hacia la información existen dos variantes de acuerdo con el tipo de tema: académico informativo o cotidiano descriptivo (Biber, 1993). Si el tema es académico, los elementos léxicos empleados podrán ser más largos, como en una presentación escolar. En este caso, se espera el uso de nominalizaciones para representar pensamiento abstracto. Por el contrario, los elementos léxicos serán cortos si la tarea solicitada es, por ejemplo, la descripción de una imagen en que se busca representar puntos más concretos. En ambos casos serán frecuentes las frases preposicionales, en el caso del académico, para subordinaciones, como en (3); en lo cotidiano, para referenciar la ubicación de elementos, como en (4). Los procesos de personalización en el lenguaje académico harán más frecuente el uso de la voz pasiva en este tipo de texto, mientras que en el plano cotidiano se espera un mayor empleo de la voz activa.

(3) *In spite of their centrality to Lakoff's theory these marginal figures have been frequently if not entirely overlooked in subsequent discussions of her work*
'A pesar de su centralidad en la teoría de Lakoff, estas cifras marginales se han ignorado con frecuencia, aunque no por completo en las discusiones posteriores sobre su trabajo'

(4) *The cat is under the table*
'El gato está debajo de la mesa'

La cuarta clase textual es la narrativa, que engloba el contar datos biográficos, experiencias de vida, o historias ficticias que pueden recurrir a una serie de imágenes como soporte. Las tareas narrativas requieren del tiempo pasado, pero los verbos no están restringidos en términos semánticos, y pueden ser materiales, mentales

o relacionales de acuerdo con las necesidades de la historia, por ejemplo, *happened*, *lived*, *thought*, *been*, *became*. Asimismo, predomina el uso de la tercera persona, como en *Snow White lived* o *Judith wrote*, y el reporte de diálogo con verbos de comunicación en cláusulas relativas con o sin *that*, como se presenta en (5) y (6).

(5) *Angela said that there were too many*
'Angela dijo que eran demasiados'

(6) *the man talked about fixing the window*
'el hombre habló sobre las reparaciones de la ventana'

La representatividad en un corpus también está determinada en términos sociolingüísticos, pues la población define el tipo de producción de lengua contenida en los textos de la colección (Baker *et al.*, 2006). Para Egbert, Biber y Gray (2022), la población y el criterio de selección definen la parte de la realidad que se desea representar, un cierto dominio de uso de la lengua. Los autores comentan que en el caso de los corpus de aprendices el criterio de selección con frecuencia obedece a una autoselección de los participantes, lo que podría poner en tela de juicio su representatividad. Es decir, los participantes en el corpus son aquellos que se sienten más cómodos con su competencia en la lengua, por lo que se sesga la muestra hacia alumnos más competentes. Esta tendencia es aún más marcada en el caso de las recolecciones longitudinales que requieren un seguimiento de al menos dos años por cada participante, ya que solo siguen participando los aprendientes más motivados.

Además de las variables sociolingüísticas que deben reportarse y considerarse para el balance de la muestra (como edad, sexo, nivel educativo, etc.), existe una manera específica y crucial de clasificar las características de los aprendices a incluir, y que incide sustancialmente en las producciones que derivan de las tareas solicitadas: el nivel de dominio de la lengua. Los diferentes niveles de dominio de la lengua inglesa, así como las funciones comunicativas orales características de estos niveles, están determinados por

el Marco Común Europeo de Referencia para las lenguas (MCER). Este marco describe la habilidad lingüística en general como un conjunto de competencias, actividades y estrategias. Así, considera esta habilidad o eficiencia como un sistema basado en tres dimensiones: competencias generales, competencias comunicativas y actividades / estrategias comunicativas. Las competencias generales de acuerdo con el MCER incluyen el saber ser, saber hacer y saber aprender; las competencias comunicativas abarcan la sociolingüística, lingüística y pragmática; y las actividades / estrategias comunicativas engloban la recepción, producción, interacción y mediación (Council of Europe, 2018).

Ahora bien, la producción oral se encuentra dentro de las actividades comunicativas junto con la producción escrita. La producción oral consta de cinco modalidades o tipos de discurso oral: monólogo descriptivo, monólogo informativo, monólogo argumentativo, discurso público y dirigirse a una audiencia. Dentro del monólogo descriptivo se encuentran las tareas que implican la narración y la descripción de información cotidiana, experiencias pasadas y expectativas a futuro, además de descripciones de situaciones y temas complejos dentro del campo de acción y/o interés del aprendiz. En este caso, el discurso utilizado puede variar en su extensión, desde frases y fórmulas memorizadas hasta el desarrollo de descripciones detalladas de subtemas relacionados. El monólogo informativo implica un flujo unidireccional de la información, aunque el hablante puede ser interrumpido ocasionalmente o para solicitarle que aclare, aporte mayor detalle o repita una respuesta inaudible. Este tipo de actividad incluye la descripción de un objeto, indicaciones e información personal sobre hechos comunes y la descripción de procesos especializados académicos o profesionales del hablante. El monólogo argumentativo abarca la descripción de intereses, opiniones y preferencias sobre diversos temas, desde los más comunes a los más especializados. Así, también puede englobar la fundamentación y defensa de la información mediante el uso de estrategias de énfasis, contraste y expansión de las posturas personales.

Como recomendaciones finales en la construcción de un corpus de aprendices, Granger (2002, 2008) y Gilquin (2015) sugieren documentar y reportar diversos aspectos relacionados con los participantes, su contexto de aprendizaje y los procedimientos de recolección de la producción de los aprendices, con la finalidad de ofrecer la información del modo más completo y transparente posible. Estos datos representan las diversas variables que afectan la producción de una L2: la lengua materna, el contexto de adquisición de la L2 (segunda lengua o lengua extranjera), el nivel de dominio de la L2, el tiempo y modo de contacto con la L2 objeto de análisis y el contacto con otras segundas lenguas, así como el contexto de recolección de los datos (educativo o natural, por ejemplo, en el trabajo o durante las actividades cotidianas). Finalmente, en el caso de que no sea una recolección natural, es necesario registrar datos sobre las tareas, como tipos, temas, condiciones (tiempo y preparación), materiales utilizados para estimular la producción y detalles sobre su aplicación, además de incluir información sobre los aplicadores de dichas tareas (sexo, lengua materna, conocimiento de otras lenguas y grado de familiaridad con el participante).

5. Metodología

El objetivo del corpus MexLeC es crear una base de datos extensa y representativa de la producción oral de aprendices universitarios de inglés como segunda lengua. Para lograr este objetivo las tareas para la estimulación de la producción oral deben ser diseñadas de acuerdo con los dos criterios principales en la construcción de un corpus de aprendices: la autenticidad y la representatividad.

Con el objetivo de capturar diversos "grados de naturalidad" de los datos se han seleccionado cuatro diferentes tareas, desde aquellas que permiten un final abierto, la libre selección del vocabulario y las estructuras con que el participante desee expresarse hasta aquellas basadas en imágenes o temáticas muy específicas, que resultan más restrictivas en relación con el vocabulario y las

estructuras requeridas para responder de manera adecuada a la tarea. Al mismo tiempo, se ha privilegiado el registro de datos monológicos no interactivos, que facilitan la producción del aprendiz en turnos extendidos, así como la mínima interrupción o participación del entrevistador o guía en la tarea, con la finalidad de obtener la muestra más extensa posible del habla de los aprendices. No obstante, los aplicadores tienen como instrucción conducir la aplicación de las tareas a modo de conversación o charla informal, y participar frecuentemente mediante interjecciones que muestren interés y eviten que el aprendiz se sienta juzgado o evaluado.

En consideración al criterio externo de representatividad de un corpus, se ha considerado el MCER (Council of Europe, 2018) para decidir las temáticas y el tipo de tarea, ya que dicho marco señala, por medio de los descriptores de la producción oral monológica, las funciones comunicativas y los temas específicos que el aprendiz es capaz de abordar en cada nivel. Como resultado, la variedad de tareas descritas en las secciones de monólogo descriptivo y monólogo argumentativo dieron lugar a las cuatro partes del instrumento aplicado que detallaremos más adelante. Aunado a ello, los descriptores del MCER sirvieron de apoyo para determinar cuáles tareas eran aplicables según el nivel de lengua del aprendiz. Esto permitió que los participantes hablaran con mayor espontaneidad y facilidad en cada una de las tareas asignadas, puesto que no se sienten rebasados o forzados al requerírseles esfuerzos más allá de sus capacidades de acuerdo con el momento en el que se encuentran de su adquisición o desarrollo de la lengua inglesa. Los descriptores para cada tarea seleccionada pueden observarse en el Cuadro 1.

CUADRO 1. Secciones en la entrevista MexLeC y descriptores correspondientes según el MCER (Council of Europe, 2018)

	TAREA 1. Preguntas sobre temas cotidianos
Nivel de lengua	Función comunicativa: monólogo sostenido. Descripción de experiencias
A1	Puede describirse a sí mismo/a, a qué se dedica y dónde vive.
A1	Puede describir aspectos simples de su vida diaria.

(*continuación*)

Cuadro 1. Secciones en la entrevista MexLeC y descriptores correspondientes según el mcer (Council of Europe, 2018)

	TAREA 1. Preguntas sobre temas cotidianos
Nivel de lengua	Función comunicativa: monólogo sostenido. Descripción de experiencias
A2	Puede describir aspectos diarios de su ambiente, como gente, lugares y experiencias de trabajo o estudio.
A2	Puede realizar descripciones breves y básicas de eventos y actividades.
A2	Puede describir planes y preparativos, hábitos y rutinas, actividades y experiencias personales pasadas.
A2	Puede describir a su familia, sus condiciones de vida, experiencia educativa y su trabajo actual o más reciente.
A2	Puede describir gente, lugares y posesiones.
B1	Puede brindar descripciones precisas sobre varias temáticas conocidas dentro de su campo de interés.
B1	Puede brindar descripciones detalladas sobre experiencias, sentimientos y reacciones.
B1	Puede describir sueños, expectativas y ambiciones.
B2	Puede brindar descripciones claras y detalladas sobre una gran variedad de temáticas relacionadas con su campo de interés.
B2	Puede describir detalladamente la importancia personal de eventos y experiencias.
C1	Puede dar descripciones claras y detalladas sobre temas complejos.
	TAREA 2. Preguntas de opción sobre preferencias e intereses
Nivel de lengua	Función comunicativa: monólogo sostenido. Descripción de experiencias.
A2	Puede explicar lo que le gusta o disgusta sobre algo.
A2	Puede usar lenguaje simple y descriptivo para comparar objetos y posesiones usando enunciados breves.
B1	Puede expresar claramente sentimientos acerca de una experiencia y expresar las razones para explicar estos sentimientos.
Nivel de lengua	Función comunicativa: monólogo sostenido. Argumentación.
A2	Puede explicar lo que le gusta y no le gusta acerca de algo. Puede explicar las razones de sus preferencias de una cosa sobre otra.
A2	Puede brindar su opinión en términos simples.
B1	Puede brindar razones simples para justificar su punto de vista sobre un tema conocido o común.
B1	Puede expresar opiniones sobre temáticas relacionadas con la vida cotidiana.
B1	Puede brindar brevemente razones y explicaciones de sus opiniones.

(*cont.*)

Cuadro 1. Secciones en la entrevista MexLeC y descriptores correspondientes según el MCER (Council of Europe, 2018)

	TAREA 3. Narrativa con base en imágenes
Nivel de lengua	Función comunicativa: Monólogo sostenido. Descripción de experiencias.
A2	Puede contar una historia o describir algo de manera general, utilizando un listado simple de hechos.
B1	Puede relatar una narrativa o descripción simple, utilizando una secuencia lineal de hechos o aspectos.
B1	Puede describir eventos reales o imaginarios.
B1	Puede narrar una historia.
C1	Puede brindar descripciones y narrativas muy elaboradas, integrando subtemas y desarrollando a profundidad hechos y detalles particulares, cerrando con una conclusión apropiada.
	TAREA 4. Preguntas de opinión
Nivel de lengua	Función comunicativa: Monólogo sostenido. Argumentación.
B1	Puede desarrollar un argumento adecuado y comprensible.
B2	Puede desarrollar un argumento de manera sistemática, resaltar puntos importantes y significativos brindando detalles relevantes que apoyen su argumento.
B2	Puede desarrollar un argumento claro y expandir y apoyar sus puntos de vista con cierta amplitud, así como brindar argumentos adicionales y ejemplos relevantes.
B2	Puede construir un argumento razonable y coherente.
B2	Puede explicar un punto de vista sobre una temática determinada presentando las ventajas y desventajas entre varias opciones.
C1	Puede sostener su postura sobre un tema complejo formulando puntos precisos y empleando efectivamente estrategias para enfatizar sus argumentos.

Con el objetivo de considerar la representatividad interna se han tratado de representar tres de los tipos textuales en Biber (2004), el discurso informativo, el discurso orientado al posicionamiento y la narrativa. Dado que el discurso orientado a la interacción constituye una tarea de tipo monológico no se ha incluido en el instrumento de recopilación. Para asegurar el balance en la producción, se han determinado tiempos mínimos y máximos para cada una de las tareas, de esta manera todos los participantes tendrían los mismos intervalos para responder, cuestión que homogeneiza la aplicación y permite la comparabilidad de los datos en la investigación.

Ahora bien, una vez determinadas la variedad de tareas (géneros textuales), las temáticas adecuadas para cada tarea y nivel de lengua (población), así como los tipos textuales por abarcar en cada una de ellas (tipos textuales) y el balance de los textos que se incluirán (tiempo para cada tarea), se pudo finalizar el guion de la entrevista para la recolección de datos. Se presenta a continuación la estructura del instrumento y los detalles de cada tarea.

6. Resultados: estructura de la entrevista MexLeC

El instrumento diseñado para la construcción del corpus MexLeC consiste en una entrevista que contiene cuatro tareas o secciones. La entrevista tiene una duración total de aplicación de 12–16 minutos dependiendo del nivel del participante. Cada una de las secciones tiene un nivel de lengua recomendado según las funciones comunicativas que el aprendiz es capaz de cumplir de acuerdo con su nivel de lengua. Además, cada tarea corresponde a una tipología textual. En el Cuadro 2 se puede observar la estructura de cada una de las partes de esta entrevista y sus detalles.

Cuadro 2. Tareas de la entrevista MexLeC

Género textual (MCER)	Tipología textual (Biber, 2014)	Descripción de la tarea	Tiempo	Nivel de lengua
Monólogo descriptivo	Informativo	Batería de preguntas relacionadas con temas como familia, amigos o tiempo libre.	3–5 minutos	Pre A1
Monólogo argumentativo	Posicionamiento	Batería de preguntas sobre preferencias o elecciones de la vida cotidiana (gastar o ahorrar el dinero, ropa cómoda o ropa de moda).	2–3 minutos	Pre A1
Monólogo descriptivo	Narrativo	Narrativa con base en una secuencia de imágenes (tira cómica a color).	3–5 minutos	A2
Monólogo argumentativo	Posicionamiento	Batería de preguntas de opinión sobre temas relacionados con el uso de la tecnología, la educación y los problemas sociales (por ejemplo, la tecnología contribuye o interfiere en la comunicación).	2–3 minutos	B1

La primera sección "Questions on familiar topics" (preguntas sobre temas conocidos) consiste en una serie de preguntas descriptivas relacionadas con la cotidianeidad del participante. Estas giran en torno a dos temáticas seleccionadas de una lista de cuatro temas centrales: familia, amigos, ocupación (escuela/trabajo) y tiempo libre. Cada una de las temáticas contiene preguntas guía que no necesariamente se repiten palabra por palabra. El criterio para juzgar el empleo correcto de las preguntas es que el participante produzca discurso según la temática general seleccionada. Se espera que en esta sección la producción lingüística del participante incluya descripciones de sus intereses, actividades, experiencias, expectativas y planes futuros. Esta es la sección más básica de la entrevista, por lo que se sugiere su aplicación para todos los niveles de lengua a partir del nivel pre-A1. El tiempo de aplicación es de 3 a 5 minutos, algunos ejemplos de estas preguntas pueden observarse en (7–9).

(7) *What do you do? Do you work or study?*
'¿A qué te dedicas? ¿Trabajas o estudias?'

(8) *What are your plans when you graduate?*
'¿Cuáles son tus planes una vez que te gradúes?'

(9) *Can you share a good memory about school?*
'¿Puedes compartirme un recuerdo bonito sobre tu escuela?'

La segunda sección, "Choice questions" (preguntas de elección), incluye tres preguntas que representan una elección de la vida cotidiana. En esta se espera que el participante exprese y explique sus preferencias e intereses. La sección puede ser aplicada a partir del nivel A1 y su tiempo esperado de duración es de 2 a 3 minutos. Algunos ejemplos de las preguntas de esta sección se muestran en (10–12).

(10) *Staying at home or going out*
'¿Quedarse en casa o salir?'

(11) *Saving or spending money?*
'¿Ahorrar o gastar el dinero?'

(12) *To study or to work?*
'¿Estudiar o trabajar?'

La sección tres lleva por título "Picture-based narrative" (narrativa basada en imágenes). En esta se muestra a los participantes una tira cómica que apareció originalmente en *Ferdinand*, obra del autor danés Mikkelsen (Ortíz, 2021). Los participantes deben crear una historia con base en la secuencia de imágenes presentada. El tiempo para esta tarea es de 2 a 3 minutos y es aplicable a partir del nivel A2 de lengua. La secuencia utilizada se presenta en la Figura 2.

La última sección de la entrevista, la sección cuatro, se titula "Opinion questions" (preguntas de opinión) y consiste en un par de preguntas argumentativas seleccionadas de cuatro opciones posibles. En esta sección se espera que los participantes expresen su opinión y establezcan argumentos que apoyen puntos de vista con relación a temáticas sobre educación, uso de la tecnología y sociedad. El tiempo esperado para esta sección es de 3 a 5 minutos. Algunos ejemplos de estas preguntas se presentan en (13) y (14).

(13) *Do you think mobile devices have destroyed communication?*
'¿Crees que los dispositivos móviles han destruido la comunicación?'

(14) *Which do you think is more practical online or face-to-face activities?*
'¿Crees que es más práctico trabajar en línea o de manera presencial?'

Es importante mencionar que para asegurar la confiabilidad en la aplicación de la entrevista MexLeC, se ha creado una guía para el entrevistador, que contiene instrucciones detalladas y las preguntas

Figura 2. Secuencia de imágenes utilizada en la tercera sección de la entrevista MexLeC (Ortíz, 2021)

específicas, así como los tiempos y objetivos (funciones comunicativas) de cada tarea. Esta guía incluye, además, una presentación (diapositivas) con los apoyos visuales necesarios. Con el objetivo de preparar a los participantes para esta entrevista, se ha diseñado también una guía rápida, que tiene como finalidad dar a conocer la estructura general de la entrevista, y que esto permita a los aprendices reducir la ansiedad y el sentido de evaluación o juicio.

Como parte de las buenas prácticas de documentación en la construcción de un corpus, a fin de complementar los materiales creados para la aplicación de la entrevista se ha elaborado un formato con el perfil de los participantes. Este formato facilita recabar datos claves para el uso de un corpus de aprendientes: nivel de dominio de la lengua, exposición a la lengua meta, lengua materna del aprendiente y de sus padres, exposición a otras lenguas extranjeras y datos personales de identificación y contacto. Todos los datos recabados se anonimizan antes de ofrecerlos al público en general.

Por cuestiones prácticas, esta entrevista se aplica de manera remota, mediante videollamada y de forma voluntaria. Los participantes se autoseleccionan, sin embargo, se invita a los estudiantes de los primeros semestres, pues el objetivo del corpus es hacer un seguimiento longitudinal durante cuatro años, tiempo promedio de estancia en la facultad. Cada uno de los participantes firma

un consentimiento para el uso de sus datos y grabaciones en cada recolección anual.

7. Conclusiones

El objetivo de este trabajo ha sido presentar el diseño de los materiales que se utilizarán para la construcción de un corpus de aprendices que contiene ejemplos de la producción oral de aprendices universitarios de inglés. La recolección pretende un seguimiento longitudinal de estos estudiantes, desde que ingresan a la facultad hasta que se gradúan, durante un tiempo aproximado de cuatro años y mediante recolecciones anuales. Este tipo de colecciones puede ser de utilidad para estudiar el proceso de adquisición de una segunda lengua, así como para diseñar materiales más efectivos con base en las necesidades específicas de los aprendices.

El corpus MexLeC sigue creciendo. Debido a su naturaleza longitudinal tendrán que pasar alrededor de 5 años antes de que alcance su tamaño final. Sin embargo, podemos reportar que en este momento contiene 142 transcripciones ortográficas que han pasado por un proceso riguroso de revisión. El corpus de producción oral resultante tiene un tamaño aproximado de 110 mil tokens. Estos datos son el resultado de las dos primeras recolecciones realizadas durante los años 2021 y 2022 en la Licenciatura en Lenguas de la Universidad Autónoma del Estado de México. A principios del año 2022, se inició la aplicación de los instrumentos y el levantamiento de datos a los estudiantes de la Licenciatura en Enseñanza de la Lengua Inglesa de la Universidad Autónoma de Hidalgo, con una segunda aplicación en 2023. En ese mismo año, el proceso también se inició en la Licenciatura en Lenguas Modernas Inglés de la Universidad Autónoma de Querétaro. Actualmente, se están afinando los detalles para la primera aplicación de los instrumentos en la Licenciatura en Lengua Inglesa de la Universidad Autónoma del Estado de Quintana Roo.

Uno de los retos más grandes relacionados con el instrumento diseñado es lograr la capacitación de los entrevistadores de mane-

ra que permita conservar la homogeneidad en la aplicación de la entrevista, y asegurar mínimas interrupciones para conservar los turnos largos del entrevistado (pero sin que el entrevistador parezca desinteresado en lo que tiene que decir el participante), así como el cuidado de los tiempos para cada tarea. Los datos recabados son de acceso libre y se encuentran disponibles para su descarga en la página web de MexLeC (Flores & Moore, 2023).

Además de la planeación de recolecciones para incluir nuevos aprendices, es necesario garantizar el seguimiento longitudinal de los participantes actuales. Otra meta a mediano plazo es comenzar la aplicación de los datos en el diseño de materiales e intervenciones didácticos, así como la mejora y reestructuración de programas que consideren las características de la lengua producida por los estudiantes.

El uso de datos de libre acceso, extensos, auténticos, representativos y estandarizados de los aprendices mexicanos de inglés permitirá contar con un modelo más amplio para el diseño de materiales, mejora de programas y planeación de intervenciones, que incluya no solo factores individuales o lingüísticos, sino también análisis empíricos cuantitativos que validen nuestras intuiciones prácticas y teóricas.

8. Referencias

Baker, Paul; Hardie, Andrew, & McEnery, Tony (2006). *A glossary of corpus linguistics*. Edimburgo: Edinburgh University Press.

Biber, Douglas (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Biber, Douglas (2004). Conversation text types: A multi–dimensional analysis. En Gérald Purnelle, Cédrick Fairon & Anne Dister (Eds.), *Le poids des mots. Actes des 7es Journées internationales d'analyse statistique des données textuelles* (pp. 15–34). Lovaina la Nueva: Presses universitaires de Louvain.

Callies, Marcus, & Paquot, Magali (2015). Learner corpus research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research*, *1*(1), 1–6. https://benjamins.com/catalog/ijlcr.1.1.00edi

Capel, Annette & Sharp, Wendy (2014). *Objective First. Student´s book with answers*. 3a. ed. Edimburgo: Cambridge University Press.

Centre for English Corpus Linguistics (2020). Learner corpora around the world. *Université Catholique de Louvain (UCLouvain)*. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Council of Europe (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume.* Estrasburgo: Council of Europe.

Egbert, Jesse; Biber, Douglas, & Gray, Bethany (2022). *Designing and evaluating language corpora. A practical framework for corpus representativeness.* Cambridge: Cambridge University Press.

Flores, Ana, & Moore, Pauline (2023). *Mexican Learner Corpus*. https://sites.google.com/view/mexlec/intro

Gablasova, Dana; Brezina, Vaclav, & McEnery, Tony (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, *5*(2), 126–158.

Granger, Sylviane (2002). A bird's–eye view of learner corpus research. En Sylviane Granger, Joseph Hung & Stephanie Petch Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Ámsterdam: John Benjamins.

Granger, Sylviane (2008). Learner corpora. En Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook* (Vol. 1, pp. 259–275). Berlín: Walter de Gruyter.

Granger, Sylviane (2021). Phraseology, corpora and L2 research. En Sylviane Granger (Ed.), *Perspectives on the L2 phrasicon* (pp. 3–24). Ámsterdam: Multilingual Matters.

Granger, Sylviane; Dupont, Maïté; Meunier, Fanny; Naets, Hubert, & Paquot, Magali (2020). *The International Corpus of Learner English. Version 3*. Lovaina la Nueva: Presses universitaires de Louvain. https://dial.uclouvain.be/pr/boreal/object/boreal:229877

Gilquin, Gaëtanelle (2015). From design to collection of learner corpora. En Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (Eds.), *The*

Cambridge handbook of learner corpus research (pp. 9–34). Cambridge: Cambridge University Press.

Gilquin, Gaëtanelle (2019). Light verb constructions in spoken L2 English: An exploratory cross–sectional study. *International Journal of Learner Corpus Research*, *5*(2), 181–206. https://benjamins.com/catalog/ijlcr.18003.gil

Hood, Susan (2010). *Appraising research: Evaluation in academic writing*. Londres: Palgrave Macmillan.

Hunston, Susan (2008). Collection strategies and design decisions. En Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook* (Vol. 1, pp. 154–167). Berlín: Walter de Gruyter.

Lehmberg, Timm, & Wörner, Kai (2008). Annotation standards. En Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook* (Vol. 1, pp. 484–500). Berlín: Walter de Gruyter.

Lüdeling, Anke, & Kytö, Merja (Eds.) (2008). *Corpus linguistics. An international handbook* (Vol. 1). Berlín: Walter de Gruyter.

Meyer, Charles F. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

McCarthy, Michael, & O'Keeffe, Anne (2009). Corpora and spoken language. En Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook* (Vol. 2, pp. 1008–1023). Berlín: Walter de Gruyter.

McEnery, Tony, & Gabrielatos, Costas (2006). English corpus linguistics. En Bas Aarts & April McMahon (Eds.), *The handbook of English linguistics* (pp. 33–71). Oxford: Blackwell.

McEnery, Tony, & Hardie, Andrew (2011). *Corpus linguistics: Method, theory, and practice*. Cambridge: Cambridge University Press.

McEnery, Tony; Xiao, Richard, & Tono, Yukio (2006). *Corpus–based language studies: An advanced resource book*. Londres: Routledge.

Meunier, Fanny (2021). Introduction to learner corpus research. En Nicole Tracy Ventura & Magali Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 23–36). Londres: Routledge.

Ortíz, Federico (2021). *Ferdinand. Ven a mi mundo.* http://www.venamimundo.com/DeAquiyAlla/TirasComicas/Ferdinand.html

Sinclair, John (1996). *Preliminary recommendations on corpus typology* (Reporte técnico). Expert Advisory Group on Language Engineering Standards. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Sinclair, John (2005). Corpus and text: Basic principles. En Martin Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxford: Owbow Books.

Wichmann, Anne (2008). Speech corpora and spoken corpora. En Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook* (Vol. 1, pp. 187–206). Berlín: Walter de Gruyter.