RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 61 (2), II Sem. 2023, pp. 13-42.

ISSN 0033-698X (impresa)
ISSN 0718-4883 (on-line)
https://doi.org/10.29393/RLA61-9DRAP20009

DISEÑO Y RECOLECCIÓN DE UN CORPUS ORAL Y LONGITUDINAL DE APRENDICES DE INGLÉS, MEXICAN LEARNER CORPUS MEXLEC[1]

DESIGN AND COLLECTION OF AN ORAL AND LONGITUDINAL CORPUS OF ENGLISH LEARNERS, MEXICAN LEARNER CORPUS MEXLEC

ANA FLORES
Universidad Autónoma del Estado de México, México.
aafloresh@uaemex.mx / https:orcid.org/0000-0002-9706-1859

PAULINE MOORE
Universidad Autónoma del Estado de México, México.
pmooreh@uaemex.mx / https:orcid.org/0000-0003-0622-904X

RESUMEN

Este trabajo presenta el proceso de diseño y recopilación del corpus *Mexican Learner Corpus* (MexLeC), a la luz de las nociones principales en la construcción de un corpus de aprendices: representatividad, autenticidad y balance (Egbert, Biber y Gray, 2022; McEnery, Xiao y Tono, 2006; Biber, 2004). Adicionalmente, se presenta un estado del arte sobre las tareas utilizadas para recopilar corpus orales de aprendices y los corpus de aprendices de inglés cuya lengua materna es el español. El resultado es un corpus oral y longitudinal único en América Latina de aproximadamente 200.000 tokens y 150 entrevistas en proceso de transcripción. Estas entrevistas representan tres tipos textuales (Biber, 2004): narrativo, informativo y de posicionamiento y dos géneros textuales (Council of Europe, 2020): monólogo descriptivo y monólogo argumentativo. Las aplicaciones principales de MexLeC son la comprensión de los patrones de adquisición de lengua extranjera, así como el desarrollo de materiales didácticos y programas de enseñanza del inglés.

Palabras clave: Lingüística de corpus; corpus de aprendices; inglés como L2; corpus orales.

ABSTRACT

This article introduces the process of design and collection of the Mexican Learner Corpus

[1]Este trabajo forma parte del proyecto "Corpus Mexicano de Aprendices" Mexican Learner Corpus" (Cód. CVU 482854), financiado por el Consejo Nacional de Humanidades Ciencia y Tecnología, México.

(MexLeC), considering the notions of representativeness, authenticity (naturalness) and balance in building a learner corpus (Egbert, Biber y Gray, 2022; McEnery, Xiao y Tono, 2006; Biber, 2004). At the same time, it presents a literature review of materials and task type in learner spoken corpora and the availability or English learner corpora from Spanish native speakers. The result is an oral and longitudinal corpus unique in Latin America of approximately 200 000 tokens and 150 interviews in transcription. The recorded interviews sample three text-types in Biber (2004), informative, stance and narrative as well as two text genres (Council of Europe, 2020), descriptive monologue and argumentative monologue. The main applications of MexLeC are in the understanding of patterns of language acquisition and the development of teaching materials and program design for English as a Foreign Language.

Keywords: Corpus linguistics; learner corpora; L2 English; spoken corpora.

Recibido: 03/05/2023. Aceptado:01/11/2023.

1. INTRODUCCIÓN

Una alternativa para complementar los estudios en adquisición, enseñanza y aprendizaje de lenguas es el uso de corpus lingüísticos de aprendices, colecciones de datos que contienen textos producidos por estudiantes de segundas lenguas en forma oral o escrita. El estudio de la lengua utilizando este tipo de corpus es un área relativamente nueva, naciente apenas a finales de los años ochenta con la creación del *International Corpus of Learner English* [ICLE] (que puede ser traducida como "Corpus Internacional de Aprendices de la Lengua Inglesa") de la Université Catholique de Louvain (Granger, Dupont, Meunier, Naets y Paquot, 2020). Esta metodología innovadora, al centrarse en la producción real de los aprendices, permite obtener una imagen más precisa de las secuencias de adquisición y de las dificultades a las que se enfrentan los aprendices de una segunda lengua (Granger, 2002). Esta colección provee una base de datos cuantificables y estandarizados que permiten validar empíricamente los resultados de estudios previos (Granger, 2008).

En el entorno mundial existen alrededor de 140 corpus de aprendices de inglés. Sin embargo, los corpus que exploran la variable "español como L1" son apenas seis: cuatro ibéricos y dos latinoamericanos. Específicamente, la variante "español mexicano" como lengua materna está contenida, no exclusivamente y solo en un pequeño porcentaje, en tres colecciones: el *Trinity Lancaster Corpus* [TLC] o "Corpus Trinity-Lancaster" (Gablasova, Brezina y McEnery, 2019), el *Longman Learners' Corpus* [LLC] o "Corpus de Aprendices de Longman" (Pearson, 2023) y el *Cambridge Learner Corpus* [CLC], "Corpus de Aprendices de Cambridge" (Cambridge University Press, 2023), siendo únicamente el primero de ellos de acceso abierto a la comunidad de investigadores en segundas lenguas.

Una observación adicional a la lista de corpus existentes es la disparidad entre el número de corpus escritos y orales, siendo los corpus escritos los que conforman la gran mayoría. Esta disparidad podría explicarse por el hecho de que la recolección de un corpus oral implica una mayor inversión de tiempo y esfuerzo. Es por ello que este tipo de corpus tienden a ser no solo muchos menos, sino también mucho más pequeños que los corpus escritos (McCarthy and O'Keeffe, 2009) y aunque parecen ser una tendencia actual en la creación de nuevos corpus de aprendientes, los corpus orales aún se consideran una gran minoría (Gilquin y Meunier 2015). No obstante, es importante resaltar que la razón principal para preferir la producción oral, en el caso de los corpus de aprendices, es la espontaneidad de la lengua hablada, pues al producirse en tiempo real, implica un grado más alto de autenticidad o naturalidad al no permitir la planeación, revisión o corrección del discurso.

En el entorno nacional mexicano, aunque los estudios sobre adquisición basados en colecciones de datos de aprendices han ido paulatinamente ganando terreno, la falta de datos cuantiosos y representativos de la variante español mexicano como lengua materna, obligan al investigador en lenguas a crear sus propias colecciones de oportunidad, que son colecciones pequeñas y poco representativas (McEnery y Hardie, 2012), a sacrificar la profundidad de su análisis, o bien, a delimitar significativamente sus preguntas y/o alcances de investigación por la inversión de tiempo necesaria en la recolección de datos.

El proyecto MexLeC parte de la necesidad de contar con una base de datos nacional y de acceso abierto, que facilite el trabajo de investigación en enseñanza y aprendizaje de lenguas basado en datos representativos de aprendices del idioma inglés como segunda lengua en el contexto mexicano. Al mismo tiempo, se espera que este corpus permita al investigador ahorrarse el tiempo de recolección que podrá invertir en indagar con mayor profundidad en sus cuestionamientos o en implementar diversas perspectivas de análisis o metodologías complementarias toda vez que el uso de datos existentes representa un ahorro considerable de tiempo y esfuerzo, siempre y cuando se cuente con información sobre el origen de los datos, su codificación y la manera en la que están organizados (Mackey y Gass, 2005). Aunado a ello, este corpus cumple con dos características particularmente útiles y relativamente escasas en las colecciones existentes: datos sobre producción oral y una recopilación de estos datos con un corte longitudinal de más de tres años.

En este trabajo se presentan los resultados de los primeros tres años de recolección del corpus MexLeC: las tareas diseñadas, los materiales utilizados, las convenciones de transcripción, los metadatos recolectados, el tipo de datos obtenidos y las posibles aplicaciones de estos en la enseñanza y aprendizaje de segundas lenguas. No sin antes, realizar un breve recorrido acerca de la lingüística de corpus de aprendices, el proceso de creación de este tipo de colecciones y las principales características de los corpus de aprendices de inglés disponibles en el mundo.

Por las consideraciones anteriores, la pregunta de investigación que sirvió de guía para la elaboración del instrumento MexLeC es ¿qué tipo de tareas solicitadas a aprendices de la lengua inglesa favorecen la recolección de una muestra de producción oral, representativa, auténtica y variada?

Corpus, lingüística de corpus y corpus de aprendices

El estudio de la lengua por medio de la lingüística de corpus implica el uso de grandes colecciones de lengua auténtica (oral o escrita) que cumplen con la finalidad de representar a cierta población y que se encuentran contenidos en formatos electrónicos (Egbert, Biber y Gray, 2022). Adicionalmente, es preciso señalar que los textos en un corpus han sido preparados para ser procesados por software específico, ya que, con este tratamiento previo, es posible analizar una gran cantidad de lengua producida utilizando métodos estadísticos para medir, contrastar y replicar los patrones encontrados en ella. El análisis con base en corpus lingüísticos permite el estudio de la lengua utilizando la lengua misma, es decir, usando la producción de los hablantes en un contexto real como evidencia empírica (McEnery y Hardie, 2012; Lehmberg y Wörner, 2008; McEnery y Gabrielatos, 2006). Esto a pesar de que ningún corpus, pese a su tamaño, puede representar la producción lingüística completa de una persona o grupo de personas.

Algunos ejemplos muy notables de este tipo de colecciones de datos son el *Corpus of Contemporary American English* [COCA] (que puede traducirse como el "Corpus de Inglés Americano Contemporáneo"; Davies, 2008) y el *British National Corpus* [BNC] o "Corpus Nacional Británico" (Brezina, Gablasova y Reichelt, 2018). El primero contiene más de un billón de palabras de la variante americana de la lengua inglesa distribuidas en diferentes géneros como ficción, revistas, periódicos, textos académicos, lengua oral, textos de la red, películas y TV desde el año 1990 hasta el año 2009. El segundo, contiene 100 millones de palabras de la variante británica de la lengua inglesa que están distribuidas en géneros como lengua oral, ficción, revistas, textos académicos y periódicos. Para el corpus BNC, incluso, ha sido diseñada una herramienta de visualización que permite analizar variantes sociolingüísticas como edad, género, región de uso, entre otras (Brezina, Gablasova y Reichelt, 2018).

En el caso del idioma español son notables los Corpus de la Real Academia Española de la Lengua, Corpus Diacrónico del Español (CORDE) y Corpus de Referencia del Español Actual (CREA). El primero consiste en una colección de 250 millones de registros que representan textos escritos de todas las épocas y lugares en los que se ha hablado el idioma español desde los inicios del idioma hasta el año 1974. El segundo, contiene más de 60 millones de formas desde 1975 hasta 2004 y contiene textos escritos de libros, periódicos y revistas además de transcrip-

ciones de radio y televisión. Es relevante también considerar la representación de la lengua española en corpus multilingües para documentar registros particulares, como es el caso del corpus *Sharing European Architecture Heritage* (SEAH) que tiene aplicaciones importantes en la enseñanza de lenguas para propósitos específicos (Colantino, 2023).

En México uno de los corpus más conocidos es el Corpus Sociolingüístico de la Ciudad de México (Butragueño y Lastra, 2012). Este corpus de hablantes mexicanos es parte del Proyecto para el Estudio Sociolingüístico del Español de España y de América, (PRESEEA, 2023) que gestiona un corpus oral que documenta las variantes geográfica y social de la lengua española. El CSCM, que documenta el español del centro de México, contiene 108 entrevistas divididas en tres niveles de instrucción: bajo, medio y alto, manteniendo una proporción 50/50 en el número de mujeres y hombres participantes.

Además de la documentación de la producción de hablantes nativos, una aplicación de gran importancia dentro de la lingüística de corpus son los corpus de aprendices; estas colecciones están construidas a partir de textos producidos por aprendientes de una segunda lengua y sus áreas de aplicación son principalmente la descripción de la interlengua, los factores que influyen en la adquisición de esta interlengua, la creación de materiales con fines pedagógicos, y el entrenamiento de herramientas de procesamiento de lenguaje natural (Meunier, 2021; Granger, 2008). Estos tipos de corpus, cuyos inicios datan de finales de los años ochenta con la creación del ICLE de la universidad de Lovaina, se han expandido a muchas otras segundas lenguas y diversos contextos de adquisición. Hasta ahora, los corpus de aprendices más notables en relación con su tamaño y quizá también con su uso en investigación y enseñanza de lenguas son: el ICLE de la Universidad de Lovaina, el TLC de la Universidad de Lancaster, el LLC de la Editorial Pearson y el CLC de la Universidad de Cambridge.

El ICLE es un corpus de lengua recopilada a partir de ensayos escritos por aprendices intermedios y avanzados. Contiene 5.5 millones de palabras y documenta el inglés escrito de aprendices de 25 diferentes lenguas maternas como árabe, búlgaro, español ibérico, chino, turco, entre otros. Su uso es únicamente con fines de investigación y bajo licencia de pago (Granger, Dupont, Meunier, Naets y Paquot, 2020). El TLC es un corpus oral, resultado de la colaboración entre el *Centre for Corpus Approaches to Social Sciences* de la Universidad de Lancaster y el cuerpo examinador de inglés como segunda lengua *Trinity College London*. Recolectado del 2012 al 2018, contiene 4.2 millones de palabras de hablantes de diversas lenguas maternas, quienes presentaron alguno de los exámenes de certificación oral Trinity GESE (Graded Examinations of Spoken English) de los niveles B1-C2 del Marco Común Europeo de Referencia sobre las Lenguas (MCER) y está disponible para su uso con fines de investigación en su herramienta web *TLC Hub* (Gablasova, Brezina y McEnery, 2019).

El CLC contiene 42 millones de palabras, que representan la producción escrita de aprendices de inglés de todos los niveles de dominio (A1-C2) que se recopilan de los diversos exámenes de certificación. Los participantes representan a más de 173 países hablantes de siete diferentes lenguas maternas. El objetivo principal de este corpus es alimentar la creación de materiales de enseñanza de la lengua inglesa, sin embargo, puede ser utilizado para investigación bajo licencia de uso o solicitud (Cambridge University Press, 2023). El LLC está conformado por ensayos y textos de exámenes, contiene al menos 10 millones de palabras y representa a todos los niveles de aprendices de inglés (A1-C2). Este corpus ha sido creado con el objeto de servir de base para la creación de diccionarios y libros para la enseñanza del idioma (Pearson Education, 2022).

El listado *Learner Corpora Around the World*, que puede traducirse como Corpus de Aprendices Alrededor del Mundo de la Universidad Católica de Lovaina (Centre for English Corpus Linguistics, 2023) guarda un registro de los corpus de aprendices recolectados en diferentes partes del mundo. En este listado se encuentran registradas aproximadamente 185 colecciones, de las cuales 166 son corpus monolingües y 19 multilingües. Los corpus en este listado documentan producciones de aprendices en 26 diferentes segundas lenguas que son: alemán, árabe, catalán, checo, chino, croata, coreano, esloveno, español, estonio, finés, francés, gaélico, holandés, húngaro, inglés, italiano, letón, lituano, noruego, persa (farsi), polaco, portugués, rumano, ruso y sueco. De igual manera documentan las lenguas maternas de los participantes, que son 22: alemán, árabe, bielorruso, catalán, chino, coreano, español, finés, hebreo, húngaro, hindi, italiano, indonesio, japonés, malayo, noruego, polaco, portugués, rumano, ruso, setsuana y sueco.

En relación con el idioma inglés como segunda lengua, se encuentran registrados aproximadamente 115 corpus, 73 de ellos escritos, 27 orales y 15 que contienen textos orales y escritos. Algunos de estos (aproximadamente 15) son colecciones multilingües, que incluyen el idioma inglés en conjunto con otras lenguas como alemán, árabe, español, francés, holandés, italiano, portugués, ruso y sueco.

Corpus de aprendices de inglés nativos del español

En el mismo listado anteriormente mencionado Corpus de Aprendices Alrededor del Mundo (Centre for English Corpus Linguistics, 2023), las producciones de aprendices que tienen el español como lengua materna consisten en apenas seis colecciones: cuatro ibéricas y dos latinoamericanas, específicamente de Chile y Ecuador. Estos seis corpus son: el *Barcelona English Language Corpus* (BELC o Corpus de la Lengua Inglesa en Barcelona) (Muñoz, 2006), el *Santiago University Learners of English Corpus* (SULEC, 2023), el *Written Corpus of Learner English* (WriCLE: O'Donnell, 2012), el *Non-native Spanish Corpus of English*, que puede traducir-

se como el corpus español de inglés no-nativo (NOSE: Díaz-Negrillo, 2012), el *English Students Oral Corpus in Chile* o Corpus Oral de estudiantes de lengua inglesa en Chile (ESOC-Chile: Zapata, 2019) y el Corpus Escrito de Aprendices de Inglés como Lengua Extranjera en Ecuador (COREAILE: Macías, 2020). De estas colecciones, cuatro contienen lengua escrita y dos incluyen producciones escritas y habladas.

El WriCLE, de la Universidad Autónoma de Madrid, está conformado por 521 ensayos de 1000 palabras en promedio cada uno, con un tamaño final aproximado de 500.000 palabras. Este corpus forma parte del proyecto TREACLE (*Teaching Resource Extraction from an Annotated Corpus of Learner English*) cuya finalidad es entender mejor la manera en la que aprenden los estudiantes universitarios españoles con el objetivo de utilizar esta información para rediseñar el currículo y crear estrategias efectivas de aprendizaje (O'Donnell, 2012). El NOSE es también un corpus escrito, conformado por aproximadamente 1000 ensayos argumentativos y descriptivos, con una extensión de entre 250.300 palabras. Sus datos han sido recolectados en un periodo de cuatro ciclos escolares en la Universidad de Granada y en un periodo de dos ciclos escolares en la universidad de Jaén con un tamaño total estimado de 300.000 palabras. Su objetivo principal es representar las dificultades de los estudiantes de estas universidades mediante un estudio profundo de la categorización en el etiquetado de errores en corpus de aprendices (Díaz-Negrillo, 2012).

El corpus BELC es de los pocos de corte longitudinal y contiene datos de 2.063 participantes: niños y jóvenes adultos que han sido recolectados en cuatro tiempos (200, 416, 726 y 826 horas de instrucción) durante un periodo aproximado de 7 años. El BELC contiene producción escrita y oral tomada de redacciones libres y de tres tareas orales consistentes en: una narrativa (usando imágenes y videos), una entrevista y un juego de roles. El objetivo del corpus es examinar los efectos de la edad en el aprendizaje del idioma inglés, por ello los grupos de recolección están formados por estudiantes cuya instrucción en la lengua inicia a los 8, 11, 14 y 18 años o más de edad (Muñoz, 2006). El grupo de 8 años cumplió los cuatro tiempos de seguimiento, el grupo de 11, tres momentos, y los grupos de 14 y 18 años o más edad únicamente completaron dos momentos de seguimiento.

El cuarto corpus español, SULEC es de la Universidad de Santiago de Compostela, iniciado en el año 2002, es un corpus oral y escrito conformado por redacciones de aproximadamente 500 palabras cada una abordando temas de opinión, además de tareas orales como entrevistas semiestructuradas, presentaciones y narrativas usando imágenes. La meta de este corpus es llegar a un millón de palabras, contando hasta el año 2022 con 400.000 tokens. Su objetivo es crear una vasta y sólida colección de lengua auténtica escrita y hablada por aprendientes de todos los niveles que sirva como base para realizar investigaciones sobre aspectos fonoló-

gicos, morfosintácticos, léxicos y discursivos en la adquisición de lenguas, además de su aplicación en la enseñanza y la traducción (SULEC, 2023).

El corpus latinoamericano COREAILE de la Universidad Técnica de Manabí, Ecuador está conformado por 210 narrativas sobre historias ficticias y personales escritas por estudiantes universitarios de nivel básico e intermedio (A1, A2 y B1) de 150-200 palabras cada una. El tamaño final de este corpus es de 44.352 palabras y tiene la finalidad de explorar la interferencia léxico-semántica y morfosintáctica del español como lengua materna en la adquisición de inglés como lengua extranjera (Macías, 2020).

Finalmente, el corpus ESOC-Chile de la Universidad Católica del Norte de Chile es una colección conformada por entrevistas semiestructuradas con una duración de 15 minutos cada una. En esta entrevista los participantes debían expresar su opinión acerca de cuatro temas principales concernientes a la realidad de Chile: el equipo nacional de fútbol, las mujeres en la sociedad, contaminación y educación chilena. En total participaron 32 estudiantes de lengua inglesa de los niveles B1 a C1 del MCER dando como resultado un total de 73.631 palabras. Este corpus está diseñado para ser utilizado por estudiantes y académicos para investigar la adquisición de la producción oral para mejorar su enseñanza (Zapata, 2019).

Además de los seis corpus mencionados, las colecciones a gran escala ICLE, TLC, CLC, y LLC también contienen producciones de aprendices cuya lengua materna es el español. Aunque, debido al gran tamaño de estas colecciones, la producción de hablantes del español representa un porcentaje mínimo y generalmente, no es posible analizar exclusivamente los datos de esta población.

Construcción de un corpus oral de aprendices: Representatividad, distribución y autenticidad

De acuerdo con Egbert, Biber y Gray (2022), un corpus es una muestra amplia de textos almacenados en formato electrónico, que son representativos de la lengua producida por una población determinada. Los corpus se diseñan bajo lineamientos específicos para identificar patrones lingüísticos en muestras representativas de producción natural. En el caso de un corpus oral, los datos son representaciones electrónicas del discurso hablado. El proceso de diseño y recolección para todo tipo de corpus debe estar encaminado a cumplir con tres características deseables: ser representativo de la producción oral de los aprendices de una L2, contener suficientes muestras para observar con precisión los patrones lingüísticos y que estas muestras sean auténticas.

Según Biber (2004), la representatividad de un corpus implica que sus muestras capturen la naturaleza variable de la lengua para que los hallazgos que se gene-

ran sobre patrones lingüísticos específicos puedan ser generalizados a la población de estudio. Para lograr esta representatividad es necesario conocer las características demográficas de la comunidad hablante y seguir un proceso estadístico para la selección de las muestras (Biber, 1993) tomando en cuenta el interés particular del investigador (Egbert, Biber y Gray, 2022). De acuerdo con McEnery, Xiao y Tono (2006), la representatividad de un corpus depende en primer lugar de la definición de la población, a lo que se denomina representatividad externa o situacional. Esta representatividad se establece en términos del ambiente social o comunicativo de uso, con atención particular en dos aspectos: la definición de las características y alcances de la población de estudio y las categorías textuales que produce esta población. La representatividad interna, en cambio, está determinada por los rasgos lingüísticos específicos de la lengua producida y la distribución de estos rasgos en la muestra y la población (Biber, 2004); es un criterio estrictamente lingüístico relacionado con las características léxicas y sintácticas de un texto.

Los corpus de aprendices son colecciones de textos orales o escritos producidos por hablantes en una lengua diferente a la materna (Granger, 2008; Meunier, 2021) por lo tanto, los conceptos clave en su diseño deben atender a las particularidades de este tipo de población y a las características de su producción lingüística.

Representatividad externa

Para establecer la representatividad externa de un corpus de aprendices, la comunidad hablante se caracteriza por variables lingüísticas demográficas muy particulares, como el historial de aprendizaje de la segunda lengua, el nivel de dominio en esta lengua, la experiencia con otras segundas lenguas, la lengua materna del aprendiz, su entorno familiar, edad, o género, entre otros.

De acuerdo con lo anterior, para lograr la representatividad en un corpus de aprendices, en primer lugar se debe identificar cuales variables demográficas serán consideradas para seleccionar a la comunidad hablante. En este caso, si la lengua objeto de recolección es la segunda, tercera o cuarta lengua; si ha sido aprendida en contextos escolares, o bien, si ha sido adquirida (y/o perfeccionada) mediante la interacción en ambientes angloparlantes. De la misma forma, deben considerarse la lengua o lenguas maternas del aprendiz y el ambiente familiar en el que ha crecido en términos lingüísticos; si es un niño, joven o adulto, hombre, mujer, etc.

Una vez seleccionadas estas variables, lo siguiente es caracterizar la lengua producida y para ello es necesario considerar dos aspectos: el nivel de dominio y los tipos de textos que existen en la lengua (géneros textuales). Con frecuencia en los corpus de aprendices el nivel de dominio de una lengua está determinado por el MCER (Council of Europe, 2020). Este marco presenta un esquema descriptivo, dividido en cuatro macrocomponentes que representan los elementos que

interactúan durante la comunicación: competencias generales, relacionadas con la identidad; competencias comunicativas, relacionadas con la lengua y su uso; actividades comunicativas, la realización de la comunicación por medio de la lengua escrita o hablada; y finalmente, las estrategias comunicativas, que son los recursos extralingüísticos usados durante la comunicación. En este caso, que se enfoca en la producción de lengua, el componente pertinente para el diseño de un corpus es el de las actividades comunicativas entre las que se encuentra nuestra área de interés: la producción (oral y escrita). En estos términos se puede clasificar a los aprendices y a los textos que producen en siete niveles, divididos en tres categorías: usuarios básicos (niveles Pre-A1, A1 y A2); usuarios independientes (B1 y B2); y usuarios competentes (C1 y C2).

En cuanto a la clasificación de los géneros textuales, en un corpus de lengua nativa esta puede atender a aspectos del medio como prensa, radio, televisión, redes sociales; al contexto de la interacción como familiar, escolar o de trabajo en un ámbito laboral específico; o bien, relacionarse con el estilo de escritura como literatura, periodístico, notas personales, etc. Sin embargo, en una población de aprendices de lengua, la producción lingüística atiende a aspectos menos naturales en ámbitos primordialmente escolares o académicos, debido a que regularmente el uso de la lengua se reduce a las actividades o tareas propuesta por un profesor de lengua o un programa académico de formación. En comparación con la lengua nativa, esta situación reduce drásticamente la variedad de texto y la posibilidad de establecer límites bien definidos entre un tipo de texto producido y otros. Una clasificación práctica de los textos que produce un aprendiz podría establecerse en relación con el tipo de actividad comunicativa y la orientación que tiene esta comunicación (Council of Europe, 2020).

De acuerdo con el MCER, las actividades de producción oral pueden clasificarse en cinco tipos: monólogo descriptivo, monólogo informativo, monólogo argumentativo, anuncios públicos y hablar en público. El monólogo descriptivo implica el uso narrativo y descriptivo de la lengua, implica la comunicación de información personal (planes, hábitos, rutinas, eventos pasados y experiencias personales) además de cualquier tema dentro del campo de interés, de estudio o trabajo del aprendiz. El monólogo informativo se caracteriza por aportar información mediante explicaciones, descripciones de objetos, eventos, o procedimientos, que pueden ser desde información cotidiana hasta procesos complejos relacionados con la vida profesional o académica. Finalmente, el monólogo argumentativo comprende la habilidad de presentar un argumento, que puede abordar desde temas simples como gustos, preferencias o intereses, hasta opiniones en torno a un tema específico y complejo. Los anuncios públicos y hablar en público se consideran géneros textuales especializados en el MCER. Los anuncios públicos abarcan actos de comunicación de información para audiencias relativamente grandes y hablar en público incluye las presentaciones en seminarios

u otros eventos de naturaleza pública (Council of Europe, 2020).

Representatividad Interna

El segundo tipo de representatividad en los corpus, la representatividad interna o lingüística está relacionada con la distribución de las estructuras gramaticales o patrones léxicos que produce la población que se pretende representar. En este sentido, el modelo multidimensional propuesto por Biber (1993) presenta un análisis factorial para identificar patrones de co-ocurrencia de rasgos lingüísticos que facilita la clasificación textual motivada por rasgos lingüísticos e independiente de factores situacionales (McEnery, Xiao y Tono, 2006; Biber, 2004).

La clasificación de los textos orales propuesta por Biber (2004) contempla cuatro tipos: informativo, de posicionamiento, interactivo y narrativo. El tipo informativo se caracteriza por el uso de nominalizaciones, adjetivos atributivos, frases preposicionales, pasivos y cláusulas relativas. El tipo dos, textos de posicionamiento, implica el uso de verbos mentales, cláusulas con "that", evasivas y adverbios de posicionamiento (probably, certainly, etc.). El tipo interactivo se enfoca en el contexto inmediato con el uso de primera y segunda persona del singular y de preguntas directas "wh". El tipo narrativo se caracteriza por el uso de verbos en tiempo pasado, pronombres de tercera persona y verbos de comunicación para reportar diálogos. Incluir tareas que fomentan la respuesta con estos tipos textuales aumenta la representatividad interna de los datos recolectados en los corpus.

Distribución

En relación con el muestreo, es necesario precisar que todo corpus es una muestra que siempre estará sesgada en algún sentido. En un corpus, el objetivo del muestreo es que este sesgo tenga un menor impacto en la colección, de tal suerte que los hallazgos cuantitativos sobre los patrones lingüísticos encontrados en el corpus se acerquen a los valores esperados en la totalidad de la lengua que representa (Egbert, Biber y Gray, 2022). Aunque no existe una colección que pueda representar de manera confiable las distribuciones de todos los patrones existentes, es importante considerar que el tipo de muestreo y el tamaño del corpus pueden acercarse al rango de variabilidad requerido. Así, un corpus demasiado pequeño no daría evidencia de fenómenos poco comunes, sin embargo, un corpus demasiado grande no siempre resulta en una mayor representación de patrones e incluso puede impactar en la practicidad y la confiabilidad de los resultados, pues en una colección demasiado grande casi cualquier análisis estadístico puede parecer significativo (Egbert, Biber y Gray, 2022). Para determinar la muestra requerida para

asegurar la precisión, existen fórmulas estadísticas con base en corpus de referencia. No obstante, en el caso de un corpus de aprendices este estimado estadístico no es de aplicación práctica, pues cada etapa de la colección tiene sus propias características y particularidades. Entonces, en este caso, la precisión del corpus puede variar a medida que se recolectan más muestras de lengua.

Ahora bien, los muestreos realizados para un corpus se describen en términos de la estratificación y la aleatoriedad (Egbert, Biber y Gray, 2022). La estratificación implica que los textos están agrupados de acuerdo con su tipo o género textual, guardando una medida proporcional para cada tipo o género textual; el muestreo proporcional, busca que la muestra represente la proporción real que existe en la lengua a través de métodos estadísticos utilizando un corpus de referencia (Egbert, Biber y Gray, 2022). En un corpus de aprendices, regularmente se prefieren las muestras estratificadas, pues se desconocen las proporciones que guarda la muestra con la población completa (Váradi, 2001). En un corpus oral, se considera que estas muestras estratificadas siempre son más representativas, pues no hay catálogos o bibliografías para consultar la distribución proporcional de los textos orales (McEnery, Xiao y Tono, 2006). Por consiguiente, en un corpus de aprendices, es preferible seleccionar una muestra estratificada y balanceada, que atienda al mayor número posible de géneros y tipos textuales, guardando un balance entre cada uno de ellos. Finalmente, la selección de los textos que representarán a cada estrato en una muestra estratificada puede hacerse de manera aleatoria o por conveniencia. En la selección aleatoria cada texto debe tener la misma probabilidad de ser elegido, en las muestras por conveniencia el investigador selecciona los textos deseados, obedeciendo a su tamaño, disponibilidad u otras características relacionadas con la practicidad. En general, para un corpus de aprendices, los textos se seleccionan por conveniencia obedeciendo, regularmente, a la disponibilidad en el contexto inmediato de los sujetos que producirán la lengua.

Autenticidad

La autenticidad de un corpus es la propiedad de capturar muestras del uso de la lengua en contextos naturales. Esto, en el caso de un corpus de aprendices, representa un concepto problemático, pues normalmente una segunda lengua aprendida en un contexto extranjero se produce como parte de las actividades de un salón de clases o en algunas pocas funciones laborales como la comunicación escrita vía correo electrónico. En este sentido, una lengua extranjera no podría ser clasificada como lengua producida naturalmente. Para la recopilación de los corpus de aprendices se considera que el grado de naturalidad del discurso puede manipularse por el control sobre la tarea y la libertad que el aprendiz ejerce lingüísticamente en la realización de las diferentes tareas (Meunier, 2021; Gilquin y Meunier, 2015;

Granger, 2008). Hablar de sus planes a futuro, por ejemplo, implica un mayor grado de libertad que describir una fotografía o repetir un diálogo ensayado donde el contenido está más controlado. Aunque, en realidad, las tres tareas requieren cierto tipo de vocabulario y estructuras sintácticas específicas, la primera tarea es mucho más abierta, porque permite al participante mayor libertad en la selección de las construcciones que utilizará para expresarse, evitando aquellas que aún no domina o prefiriendo aquellas que le resultan de fácil acceso.

Constitución de los corpus existentes: Tareas y materiales

Los corpus orales disponibles en el listado *Learner corpora around the world* (Centre for English Corpus Linguistics, 2023) utilizan diversas tareas que van desde la más controladas como repeticiones o lectura en voz alta, hasta las que permiten una producción más espontánea, sin preparación previa y con finales abiertos. Estas podrían clasificarse en cuanto al modo discursivo: en tareas monológicas y de interacción; en cuanto al grado de control ejercido por el diseñador de la tarea: de controlado a libre; y por el grado de espontaneidad esperado del participante: de planeada a espontánea (Council of Europe, 2020; Mackey y Gass, 2005).

Las tareas monológicas son aquellas que implican turnos extendidos por parte de los participantes, regularmente sin sufrir interrupciones, cuyo objetivo es permitir al aprendiz expresar de manera libre e ininterrumpida experiencias, puntos de vista, opiniones e información personal y/o factual solicitada. Se favorecen este tipo en el diseño de instrumentos que pretenden recopilar corpus de aprendices, debido a que la meta es recuperar la mayor cantidad posible de uso de la lengua de los participantes. Las tareas comprendidas en esta clasificación son cinco: las narrativas, las descripciones, las preguntas de opinión, las presentaciones, y las repeticiones textuales.

En las tareas narrativas los aprendices deben contar una historia original con base en su experiencia, relacionado a una secuencia de imágenes o pautas textuales asignada, o bien recontar una historia conocida por la comunidad. Los materiales para este tipo de tareas suelen ser imágenes, videos, lecturas, títulos de historias famosas o preguntas específicas para evocar experiencias personales. Las imágenes pueden ser lineales, en blanco y negro o realistas, y representar a los personajes o eventos clave sobre las que el participante debe elaborar una historia en el espacio de uno a tres minutos. Algunos corpus de aprendices emplean formatos creativos para solicitar la narrativa. Es el caso del *Giessen-Long Beach Chaplin Corpus* (GLB-CC) de la Universidad de Giessen, Alemania en el que los aprendices participan en parejas y observan extractos diferentes de una película de Charles Chaplin para contar el uno al otro la secuencia respectiva (Jucker, Müller y Smith, 2006). También en el *Corpus Parallèle Oral en Langue Étrangère* o Corpus Paralelo en Lengua

Extranjera (PAROLE) de la Universidad de Savoie, Francia los participantes deben contar un accidente que han sufrido. Para ayudarles a recordar algún incidente se les muestran algunas imágenes: una persona con el brazo vendado, una ambulancia, un auto descompuesto y una persona utilizando muletas (Hilton, 2008).

Las tareas de descripción con frecuencia están basadas en fotografías. Se muestra a los participantes una fotografía para que describan las acciones, características y las actividades retratadas. En el corpus COREIL de la Universidad de París-Diderot se utilizan dos tipos de imágenes: (1) una imagen estática, donde las personas no realizan acción alguna y (2) una dinámica, en donde se representan personas realizando actividades, lo que promueve el uso de los distintos patrones gramaticales del interés de los creadores del corpus (Delais-Roussarie y Yoo, 2011).

Las tareas monológicas argumentativas, pueden incluir las preguntas de opinión, representan temáticas específicas y relativamente complejas (no relacionadas con la vida cotidiana) que implican un turno extendido en el que el aprendiz expresa su opinión sobre la temática asignada, además de formular argumentos que sustenten su postura. En el *International Corpus Network of Asian Learners of English*, que puede ser traducido como la Red Internacional de Corpus de Aprendices Asiáticos de Lengua Inglesa (ICNALE), los participantes deben hablar durante 60 segundos sobre algunas temáticas específicas como "Part-time jobs for students" o "Smoking banned at restaurants" (Ishikawa, 2023). En el corpus ESOC-Chile (Zapata, 2019), los participantes deben elegir uno de cuatro temas, por ejemplo, "La educación en Chile" y "El rol de las mujeres en la sociedad chilena".

Algunos corpus ocupan tareas para recopilar el "habla en público" identificado por el MCER como un discurso especializado (Council of Europe, 2020). Estas tareas normalmente consisten en presentaciones donde se facilita la manipulación de la espontaneidad según se le proporciona tiempo para planear la intervención o no al participante. Los exámenes orales de Trinity usados para recopilar el TLC contempla una presentación previamente planeada por el candidato. Otros aspectos de la tarea abiertos a la manipulación en el diseño son la selección del tema (previamente designado o seleccionado por el participante) y el acceso a materiales de referencia y apoyo durante la presentación. En el *International Teaching Assistants Corpus* o Corpus de Asistentes Internacionales de Enseñanza (ITAcorp) de la Universidad del Estado de Pensilvania, se recopilaron las presentaciones que realizaban los estudiantes como parte de un curso de preparación para asistentes de enseñanza. En el corpus ANGLISH de la Universidad de Provenza (Francia), el participante entra a una cabina de grabación y se le pide hablar sobre el tema de su elección durante dos minutos. Se le proporcionan dos minutos previos para identificar y preparar este tema (Tortel, 2008).

El mayor grado de control y menor nivel de autenticidad se observa en el último grupo de tareas monológicas, las de repetición, utilizadas para evaluar aspectos

fonológicos y fonéticos. Estas consisten en la lectura o repetición de textos asignados para medir aspectos específicos de pronunciación y entonación mediante la lectura de listas de palabras, oraciones, o textos breves en forma de narrativas o diálogos. En el corpus EVA, por ejemplo, los participantes leen en voz alta una conversación adoptando un rol cada uno (Hasselgren, SF). En el corpus EUROM1, una base de datos diseñado para comparar las características fonológicas de siete lenguas europeas, los materiales consisten en cuatro listados para lectura en voz alta: monosílabos, extractos textuales, enunciados y números seleccionados de manera sistemática para reflejar diversos elementos fonológicos (University College London, 2023). En el corpus ICE-IPAC (*Phonologie de l'Anglais Contemporain*) los participantes deben repetir un listado de palabras que contiene todos los fonemas de la lengua inglesa. Los materiales se dividen en dos listas: una genérica para todos independiente de su lengua materna y una específica de acuerdo con la L1 del participante (Kamiyama, Lacoste y Herry-Benit, 2017).

En el segundo grupo, las tareas de interacción, en lugar de un turno extendido del aprendiz la producción se genera en interacción con un interlocutor quien puede ser un compañero o un entrevistador. Estas tareas introducen cierta complejidad metodológica por el juego de poder y/o solidaridad entre los participantes que pueden afectar el tono de la interacción. Regularmente, se presentan como parte de una entrevista semidirigida en la que se intenta generar un intercambio orgánico, más auténtico y espontáneo. En los corpus analizados se encuentran las siguientes actividades interactivas: conversaciones, discusiones, juegos de roles e interacciones áulicas.

En las conversaciones, los participantes deben abordar un tema general con otro participante o con el entrevistador en el que se privilegia el intercambio de opiniones, puntos de vista o información que giran en torno a un tema específico. Las temáticas utilizadas en las conversaciones informales están relacionadas con la vida cotidiana o estudiantil, las ocupaciones, preferencias e intereses, pasatiempos, algunas experiencias pasadas importantes, planes y sueños para el futuro. Sin embargo, existen excepciones notorias, como el *Tübingen Corpus of Eastern European English* (TCEEE) de la Universidad de Tübingen, Alemania que incorpora temas más especializados como la historia de aprendizaje, la vida profesional y el involucramiento de los participantes en proyectos internacionales (Salakhyan, 2012).

Las discusiones giran en torno a la solución de problemas. Los participantes interactúan en parejas y deben tomar una decisión consensuada sobre situaciones específicas. Es común que, dentro de los materiales para este tipo de tarea, se brinde a los aprendices un diagrama o serie de ilustraciones que les sirva de inspiración, ofreciendo alternativas para la solución del problema, como es el caso en los exámenes de certificación utilizados para recopilar el corpus CLC. Una excepción es el procedimiento en el corpus GLBCC donde la discusión está basada en los

sucesos de una película, que los participantes ven para después narrar su extracto a su compañero (Jucker, Müller y Smith, 2006).

El juego de roles es un instrumento en el que se especifican papeles, características, o posturas para los participantes. Una variante interesante de juego de roles es la adoptada por el corpus ICNALE, de Kobe University, llamada juego de roles persuasivo. En esta tarea el participante asume un papel en el cual debe convencer a otro estudiante a realizar una acción que no desea realizar o con la cual no está conforme. Para esto, hay dos situaciones que el participante debe enfrentar: en la primera el participante es un estudiante con un trabajo de medio tiempo y debe persuadir a su supervisor de conservar su trabajo, mientras el otro tiene la instrucción que los estudiantes no deben trabajar. En una segunda situación, un participante adopta el rol de un comensal en un restaurante que exige la devolución de su dinero, pues no le ha sido posible disfrutar de sus alimentos debido a la gran cantidad de gente fumando en este lugar. Este tipo de tarea permite alejar a los participantes de los roles ensayados, permitiendo evaluar la competencia comunicativa pragmática en la lengua meta además de suponer un mayor riesgo e involucramiento cognitivo en la tarea (Ishikawa, 2023). Otro juego de roles no ensayado se realiza en el corpus *Finnish Upper Secondary School Corpus of Spoken English* o Corpus Finés de Inglés Hablado en Escuelas Secundarias (FUSE) de la Universidad de Helsinki, en el que se pide a los participantes actuar como si fueran dos amigos que no se han visto por mucho tiempo (Ehrnrooth, 2015).

Finalmente, algunos corpus se integran por interacciones dentro de un contexto real. En este caso, la mayoría se ubican en contextos escolares, documentando así no solo la lengua de los estudiantes, sino también la producida por los profesores o los materiales de clase como libros de texto, exámenes y tareas. En este caso por la particularidad de las situaciones que documentan resaltan los corpus *Michigan Corpus of Academic Spoken English*, que puede traducirse como Corpus de Inglés Académico Hablado de Michigan (MICASE) de la Universidad de Michigan (Simpson, Briggs, Ovens y Swales, 2002) y la *European Science Foundation Second Language Database* (Base de Datos de Segunda Lengua de la Fundación Europea de Ciencias) del Instituto Max Planck (Feldweg, 1992). En el primer caso, el corpus recoge interacciones producto de eventos académicos en la Universidad de Michigan con el objetivo de documentar el discurso académico de los aprendices de inglés. Los eventos incluyen clases, conferencias, coloquios, presentaciones de estudiantes, seminarios, sesiones de laboratorio, interacciones en oficina, grupos de estudio, reuniones, asesorías, tutorías individuales, entrevistas, recorridos guiados en el museo y defensas de trabajos de grado, entre otros (Simpson, Briggs, Ovens y Swales, 2002). En el caso de la Base de datos ESF los textos recolectados consisten en lenguaje espontáneo producido por 40 trabajadores inmigrantes del occidente de Europa, que interactúan con hablantes nativos en sus respectivos países de acogida (Feldweg, 1992). El grado de autenticidad en estos corpus es alto,

pero también la recolección de datos genera altos costos.

Se han descrito las tareas de recopilación aquí de manera aislada, sin embargo, la mayoría de los instrumentos de colección de datos para corpus incorporan una selección de tareas para asegurar la representatividad interna por solicitar diversos tipos y géneros textuales.

2. METODOLOGÍA

El objetivo del corpus MexLec es crear una muestra representativa de la producción oral de los aprendices mexicanos de inglés, que permita observar el proceso de adquisición de esta lengua, por medio del seguimiento de estos aprendices a través de los tres a cinco años que permanecen en formación universitaria como docentes o traductores de inglés.

Participantes. La población objeto son estudiantes universitarios especializándose en lengua inglesa como docentes, traductores o intérpretes de inglés. Estos estudiantes permanecen en las instituciones de recolección de cuatro a cinco años divididos en periodos semestrales. Hasta ahora, los participantes son 78 estudiantes, 22 hombres y 46 mujeres entre 18 y 25 años de la Universidad Autónoma del Estado de México. Sus niveles de dominio de la lengua son de A1-B1 dependiendo del semestre de recolección. La lengua materna de estos estudiantes es español mexicano; esta lengua es también la lengua de convivencia con sus padres, aunque uno de los participantes ha mencionado que la lengua materna de uno de sus padres es el idioma alemán. El aprendizaje de la lengua meta (inglés) de la mayoría de los participantes ha sido por medios escolarizados, como parte de sus estudios básicos y medio superior que comprende desde año y medio hasta diez años de instrucción previa al ingreso a la universidad. Para la primera recolección (2021) nueve de los participantes mencionan haber visitado un país de habla inglesa, número que se conserva para la segunda recolección (2022). Finalmente, la mayoría de los participantes mencionan tener contacto con la lengua francesa y seis más mencionaron que también tienen contacto con otras lenguas como japonés, italiano y alemán.

Piloteo del instrumento. La recolección piloto se llevó a cabo durante el año 2020, con estudiantes de nuevo ingreso y de cuarto año de la licenciatura en lenguas. Un total de 37 entrevistas, 7 hombres y 30 mujeres con un rango de edades de los 18 a los 36 años, todos ellos hablantes de español mexicano como lengua materna y familiar. Este piloteo sirvió además de hacer ajustes al instrumento y materiales, para pilotear el procedimiento y las convenciones de transcripción. De este piloteo resultó el ajuste de las tareas de tres a cinco minutos cada una de ellas,

un intervalo práctico que permite a los participantes de niveles inferiores lograr un tiempo mínimo y que a su vez ofrece suficiente tiempo para que los aprendices avanzados terminen su discurso. En el instrumento para el perfil del estudiante, el piloteo permitió precisar las preguntas, además de agregar algunas preguntas sobre dominio de la segunda lengua. Los materiales de aplicación se adaptaron para minimizar el impacto del discurso del entrevistador en la producción del entrevistado. Específicamente, se incorporó una presentación con láminas para reducir la interacción con el entrevistador evitando la toma de turnos largos y preguntas de contenido que pudieran guiar el discurso del aprendiz y generar problemas de comparabilidad entre las entrevistas. Sin embargo, el entrevistador guarda la función de mostrar interés y expresar mediante monosílabos y frases cortas la instrucción de continuar o agregar información.

Diseño final: instrumento de recolección. Para el diseño de las tareas y la recopilación de las muestras de lengua, se han considerado dos de las actividades comunicativas monológicas descritas en el MCER (Council of Europe, 2020): monólogo descriptivo y monólogo argumentativo y tres de las tipologías textuales orales en Biber (2004): textos informativos, textos narrativos y textos de posicionamiento. Dando como resultado el diseño de cuatro tareas representando, cada una de ellas, a estos diferentes géneros y tipos textuales seleccionados, que una vez recolectados los datos formarán los estratos del corpus.

El instrumento de recolección resultante es una entrevista, que consiste en mostrar a los participantes frases o preguntas para motivar la producción oral. Esta entrevista está dividida en cuatro secciones o tareas, con una duración total de 12-20 minutos (3-5 min cada sección). En la primera sección, denominada "Questions on familiar topics", los aprendices deben hablar sobre sí mismos en relación con sus amigos, familia, ocupación y tiempo libre, un reactivo de ejemplo es una lámina mostrando la leyenda: "My friends" a partir de lo que se espera que el participante describa sus amistades. Esta primera tarea representa el género textual monólogo descriptivo, a su vez que representa el tipo textual informativo. En la segunda sección, llamada "Choice questions", los participantes deben responder preguntas que les presentan opciones, deben seleccionar una de ellas y expresar las razones de sus elecciones y preferencias, un reactivo de ejemplo es: "Study or work? Why?". Esta sección representa el monólogo argumentativo y tipo textual de posicionamiento. En la tercera sección de la entrevista, a la que se le asignó el nombre "Story-telling", los aprendices deben contar una historia utilizando una secuencia de imágenes (Ortiz, 2021). El discurso producido en esta sección busca representar el género textual monólogo descriptivo y el tipo textual de posicionamiento. Finalmente, en la cuarta sección, llamada "Opinion questions", los aprendices deben responder preguntas de opinión sobre temas como educación, sociedad, trabajo y uso de la tecnología. Un reactivo de ejemplo es "Do you think

mobile devices have destroyed communication or have they made it easier?" Esta última sección representa el género textual monólogo argumentativo, así como el tipo textual de posicionamiento. La diferencia entre esta sección y la segunda es en términos de la distancia con el tema, en la segunda sección se abordan temas familiares y en esta parte se abordan temas menos conocidos. De esta manera cada una de las cuatro tareas sirve como un estrato que guarda un balance en la recolección entre cada participante, relacionado con el tiempo asignado. Las diferentes secciones de la entrevista MexLeC y los tiempos asignados para las mismas, pueden observarse en la Figura 1.

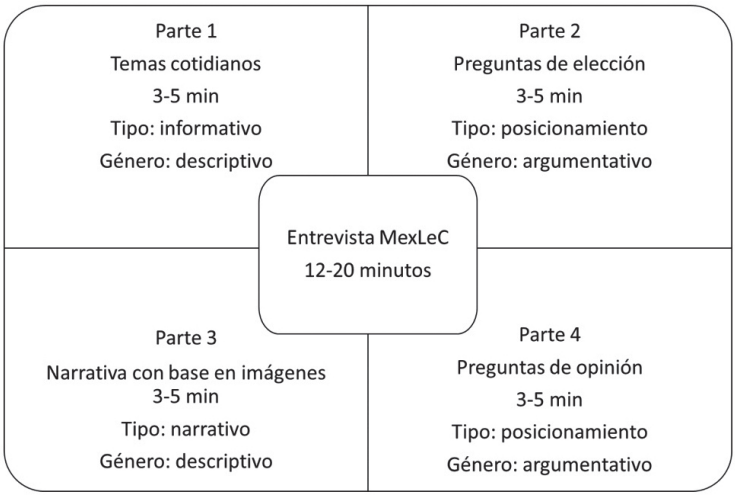

Figura 1. Tareas en la entrevista MexLeC.

Distribución y Muestreo. Para la selección de las muestras, representadas por los participantes cuya producción se ha incluido en el corpus, se realizó un muestreo por conveniencia, pues a medida que pasan los momentos de recolección o años de seguimiento (primer año, segundo, tercero, cuarto, quinto…), algunos participantes se retiran del proyecto y sus datos son eliminados de la base de datos. En este sentido, las producciones seleccionadas responden a la participación voluntaria de los aprendices y a su permanencia regular dentro de la institución. El tipo de distribución seleccionada para el corpus ha sido estratificado; guardando un balance o proporcionalidad entre los textos producidos por cada participante, relacionados con el tiempo brindado para cada tarea (3-5 minutos por tarea y 12-20 minutos en total). Este tiempo ha sido designado para capturar la cantidad de producción que el aprendiz produce de acuerdo con su nivel de dominio de

la segunda lengua, estableciendo así un límite práctico. Es importante mencionar que, en relación con los tipos textuales, la muestra presentará una desproporción en las diferentes categorías consideradas pues los tipos textuales narrativo e informativo recolectan una producción de tres a cinco minutos cada uno, mientras que los textos de posicionamiento se recolectan en dos tareas, esto es de 6-10 minutos de producción. Esta proporción se ha permitido, dado que la tarea cuatro probablemente es difícil para los participantes iniciales (A1-B1), pues de acuerdo con los descriptores del MCER, en estos niveles de lengua, los hablantes aún no están listos para este nivel de argumentación y/o vocabulario complejo.

Registro de participantes. El proceso de recolección del corpus MexLeC inicia con un registro de los datos de contacto de los aprendices. En este registro se solicita, además, información sobre la lengua materna del participante y la de sus padres, información sobre sus antecedentes en el aprendizaje y experiencias con la lengua de recolección (inglés) y el contacto con otras segundas lenguas. Una vez que se ha recolectado está información, la entrevista es aplicada por medio de videollamada con la aplicación Zoom. La llamada es videograbada y la grabación es codificada con el identificador de la universidad y el grupo de estudiantes o generación. En cuanto a los procesos administrativos, es importante mencionar que cada uno de los participantes firma un consentimiento para ser videograbado, para el uso de sus datos y su inclusión en el corpus. Asimismo, los datos de identificación se eliminan de todas las entrevistas y materiales de acceso abierto. Sin embargo, se conserva un registro anónimo de acceso abierto, del perfil de estos participantes, lo que permitirá visualizar datos de utilidad para los usuarios/investigadores durante el análisis e interpretación de los datos.

Convenciones de transcripción. Una vez codificado el archivo MP4 (audio y video) se transcribe ortográficamente el discurso utilizando un listado de convenciones adaptado del TLC (Gablasova, Brezina y McEnery, 2019) y el LINDSEI (Gilquin, De Cock y Granger, 2010). Los aspectos marcados en estos corpus son los turnos, traslapes, pausas, palabras truncadas y falsos inicios. Las palabras mayúsculas son utilizadas solo si se han deletreado y los números se escriben con letras para evitar ambigüedad, las formas contraídas se mantienen al igual que las contracciones informales y se omite la puntuación en todas las transcripciones. Se codifica, además, información léxica como palabras extranjeras e interjecciones. Así mismo, se marca información extralingüística y fonológica como calidad de la voz, alargamiento de vocales, sonidos inaudibles, información contextual y vocalizaciones que no son del habla (risas o estornudos). Finalmente, ambos corpus incluyen etiquetas que indican la tarea de la que se trata y presentan los datos sensibles anonimizados bajo etiquetas tipo “nombre” o “lugar”; y adicionalmente, en TLC se agregan etiquetas que indican los tiempos dentro de la grabación. Para

el corpus MexLeC se ha agregado la marcación de la entonación interrogativa y se han omitido aspectos como los traslapes, pues el discurso capturado es principalmente monológico; se han omitido las etiquetas de tiempo y los aspectos fonológicos relacionados con el alargamiento de vocales y la calidad de la voz.

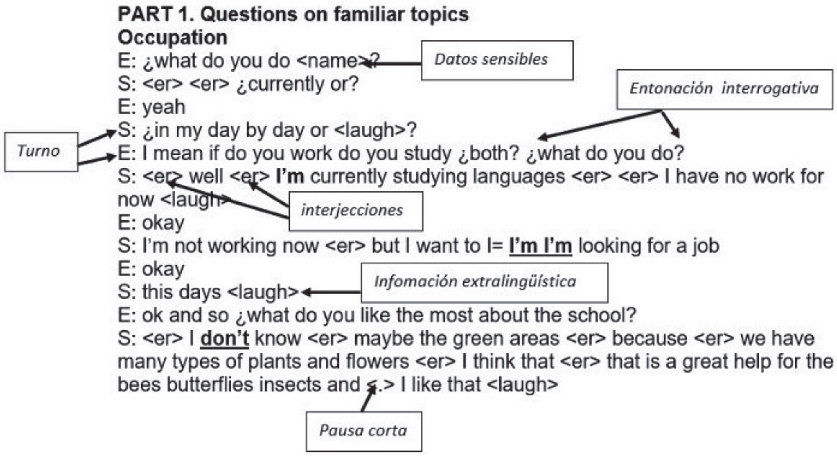

Figura 2. Extracto de entrevista señalando algunas de las convenciones adoptadas en el corpus MexLeC (Flores y Moore, 2023).

En la Figura 2 se pueden observar algunas de las convenciones adoptadas en el corpus MexLeC, como las interjecciones, datos sensibles, etc. Adicionalmente en resaltado se encuentran los datos de la sección de la entrevista/tarea y el tema; en rojo aparecen ejemplos de la transcripción de algunas contracciones.

Etiquetado. Las entrevistas transcritas se convierten a texto plano para su procesamiento y son cargadas en la página web del corpus MexLeC (Flores y Moore, 2023). Posteriormente, los archivos de texto (txt) son etiquetados semi automáticamente en dos versiones de acuerdo con el interés del investigador: utilizando TagAnt (Anthony, 2022) y CLAWS7 (UCREL 2023). TagAnt es un software de uso libre que puede realizar un etiquetado automático de textos en 24 diferentes idiomas, en el caso de la lengua inglesa, cuenta con un total de 58 categorías gramaticales básicas como "UH=interjeccion" (e.g. er) y "DT=determinante" (e.g. some), además de etiquetas relacionadas principalmente con flexiones nominales, verbales y adjetivales como "NNS=sustantivo plural" (e.g. horses), "VVD=verbo en pasado" (e.g. danced) y "JJR=adjetivo comparativo (e.g. bigger). CLAWS7 contiene 137 categorías gramaticales que incluyen todos los tipos mencionados

anteriormente con mayor detalle, como "PPHO1= pronombre personal objetivo de tercera persona singular" (e.g. him, her); además de algunas etiquetas que representan aspectos sintácticos como "DA=adposición o posposición determinante capaz de cumplir una función pronominal" (e.g. such) y semánticos como "FW=palabra extranjera" (e.g. jugar).

<_XX er_UH >_XX <_XX er_XX >_XX ¿_ADD currently_RB or_CC ?_.
¿_VBZ in_IN my_PRP$ day_NN by_IN day_NN or_CC <_NNS laugh_NN >_ADD ?_.
<_XX er_UH >_XX well_UH <_XX er_UH >_XX I_PRP 'm_VBZ currently_RB studying_VBG languages_NNS
<_XX er_UH >_XX <_XX er_SYM >_XX I_PRP have_VBP no_DT work_NN for_IN now_RB <_XX laugh_NN
>_XX
I_PRP 'm_VBP not_RB working_VBG now_RB <_XX er_UH >_XX but_CC I_PRP want_VBP to_IN I=_PRP I_PRP
'm_VBP I_PRP 'm_VBP looking_VBG for_IN a_DT job_NN
these_DT days_NNS <_XX laugh_VBP >_XX
<_XX er_UH >_XX I_PRP do_VBP n't_RB know_VB <_XX er_UH >_XX maybe_RB the_DT green_JJ areas_NNS
<_XX er_UH >_XX because_IN <_XX er_UH >_XX we_PRP have_VBP many_JJ types_NNS of_IN plants_NNS
and_CC flowers_NNS <_XX er_UH >_XX I_PRP think_VBP that_IN <_XX er_UH >_XX that_WDT is_VBZ a_DT
great_JJ help_NN for_IN the_DT bees_NNS butterflies_VBZ insects_NNS and_CC I_PRP like_VBP that_DT
<_XX laugh_NN >_XX

Figura 3. Extracto de texto plano, etiquetado.

En la Figura 3 se observa un extracto de la misma sección de la entrevista en la figura 2. En este extracto se ha eliminado el discurso del entrevistador y los turnos y ha pasado por un proceso de etiquetado automático utilizando el software TagAnt (Anthony, 2022).

3. RESULTADOS Y DISCUSIÓN

Actualmente el corpus MexLeC cuenta con aproximadamente 200.000 tokens distribuidos en dos cortes generacionales, con tres muestras (una por año) de cada uno. Estos datos han sido recolectados durante los años 2021, 2022 y 2023 en la Universidad Autónoma del Estado de México. Adicionalmente, se trabaja en dos ramas del mismo corpus en universidades públicas de los estados de Hidalgo y Querétaro, México. En la Universidad Autónoma del Estado de Hidalgo se cuentan con 23 entrevistas transcritas y listas para su uso que han sido recolectadas en el año 2022 y se trabaja en la transcripción de una segunda muestra. En la Universidad de Querétaro se encuentran en proceso de transcripción 53 entrevistas como primera muestra. El total final de entrevistas transcritas disponibles hasta ahora es de 139, con un promedio de aproximadamente 850 palabras cada una. Los niveles de estas producciones van desde A1 a B2 del MCER. Se determinaron los niveles con base en el plan de la materia de lengua inglesa, las horas y materiales de estu-

dio, así como los exámenes de colocación internos y de práctica que se aplican a los estudiantes cada semestre. Las entrevistas aplicadas han sido videograbadas cuidando la calidad del sonido y la imagen del participante. Estas videograbaciones se encuentran disponibles bajo petición para fines de investigación del lenguaje no-verbal. Las universidades incluidas en MexLeC, los años de seguimiento y el total de entrevistas por universidad, pueden observarse en la Tabla 1.

Tabla 1. Universidades y recolecciones MexLeC durante los años 2021-2023.

Universidad Autónoma	**Recolección 2021**	**Recolección 2022**	**Recolección 2023**
Estado de México (UAEMex)	SI (72 entrevistas)	SI	SI
Hidalgo (UAEH)	NO	SI (32 entrevistas)	SI
Querétaro (UAQ)	NO	NO	SI (53 entrevistas)

Los datos contenidos en el corpus MexLeC representan el discurso producido por aprendices de inglés como lengua extranjera, en un contexto universitario mexicano. Las secciones (estratos) de este corpus se dividen en cuatro: "1. Monólogo descriptivo", "1.1. tipo textual informativo", "1.2. Tipo textual narrativo" y "2. Monólogo argumentativo". Las secciones numeradas como 1 y 2 representan los géneros textuales relacionados con las funciones comunicativas que buscan cumplir los hablantes (Council of Europe, 2020). La sección uno, monólogo informativo, contiene discurso que hace descripciones de hábitos, rutinas, intereses y aspectos de la vida cotidiana del aprendiz como la familia, los amigos, la ocupación (trabajo o estudios) y las actividades de tiempo libre. Adicionalmente, esta sección incluye la narración de historias siguiendo una secuencia lógica-cronológica de hechos. La sección dos, monólogo argumentativo, incluye discurso, en el que el aprendiz expresa sus preferencias y compara opciones, además de expresar y sustentar sus opiniones sobre temas como tecnología, trabajo, educación y problemas sociales del entorno nacional.

Para representar los tipos textuales (Biber, 2004), que son los rasgos lingüísticos esperados en el discurso producido por los aprendices, se identifican tres secciones en el corpus MexLeC, que son las descritas como 1.1. Tipo textual informativo, 1.2. Tipo textual narrativo y 2. Monólogo argumentativo. El discurso en la primera sección contiene textos orientados a la información, que son producidos cuando los aprendices comparten información personal, sobre sus hábitos y rutinas y sobre hechos específicos presentes y futuros con la única intención de informar al entrevistador. El discurso contenido en la sección 1.2 son textos con

características lingüísticas narrativas sobre historias o hechos pasados. Finalmente, las producciones contenidas en 2, Textos de posicionamiento, se generan cuando los aprendices exponen razones y argumentos que sostienen un punto de vista, una preferencia o una opinión.

El grado de naturalidad de los datos en MexLeC está condicionado a la producción sobre temas específicos, que van desde los más familiares sobre la vida cotidiana, hasta temas más complejos relacionados con la sociedad, la tecnología, la educación y el trabajo. En cuanto al nivel de control del discurso producido, la tarea más condicionada, es la narrativa basada en imágenes, pues modela el discurso al seguimiento de una secuencia de imágenes específica. En las preguntas de elección y opinión el discurso producido es más espontáneo, sin embargo, aún se limita a las opciones y temáticas solicitadas y debe culminar con la consecución de la función comunicativa que es sustentar la opción elegida. Finalmente, la tarea con mayor grado de libertad es la tarea sobre temas cotidianos, en ella, aunque el tema es proporcionado por el entrevistador y se manipula la temporalidad, se permite un final abierto y tiende a permitir que el participante seleccione la información que desea intercambiar.

En una observación simple de frecuencias prevalece el uso de pausas e interjecciones, así como repeticiones y usos de la lengua materna para llenar vacíos léxicos, situaciones que a primera vista van disminuyendo a medida que se observan las producciones de aprendices con niveles más altos de dominio.

Por ejemplo, un participante de nivel A2 responde acerca de lugares que le gustaría visitar con la siguiente participación:

> *because I I love the well I I have seen a lot of interesting places in California I would like to visit I would like to go to London to see <er> I really love the <er> Harry Potter movies and I think there's iconic place I think I don't know very much the what things are there but I I know that I would like to go there* (MexLeC, UAEMex, participante 1, primer levantamiento de datos)

Se observan muchas repeticiones que son naturales en el contexto porque le proporciona tiempo para planear su discurso. También hay pausas rellenas con el marcador de duda "er". A pesar de lo accidentado que se ve la transcripción, en la oralidad transcurre de manera relativamente fluida, con frases extendidas entre las pausas y se entiende el contenido en el sentido de que ha visitado California y le gustaría ahora visitar lugares relacionados con las películas de Harry Potter en Londres.

La frecuencia de vocablos en el corpus MexLeC es similar a lo esperado para discurso hablado (Leech, Rayson y Wilson, 2001). En la Tabla 2 se puede observar las 30 palabras más frecuentes en el primer levantamiento de datos en la UAEMex:

Tabla 2. Frecuencia relativa de palabras en los datos MexLeC.

Tipo	Frecuencia	Dispersión
I	2616	0.210915
and	1735	0.355807
the	1666	0.415625
to	1276	0.48473
my	995	0.446045
a	804	0.450005
is	732	0.6458
in	714	0.596451
like	684	0.5521
that	590	0.797604
with	534	0.518935
because	524	0.474069
don't	512	0.51536
we	473	0.769661
but	465	0.51263
have	427	0.54569
think	416	0.786442
it	415	0.720509
so	364	0.842971
of	344	0.799022
you	342	0.910089
know	339	0.796
yes	321	0.72628
well	315	0.978133
was	313	0.734322
for	313	0.870891
or	298	0.727293
he	292	0.711446
his	287	0.719296
me	232	0.840444

Esta tabla presenta la palabra tipo no lematizado, es decir, no separa *'like'* en su uso verbal de *'like'* en su uso preposicional o como marcador discursivo. En la se-

gunda columna se encuentra el dato del número de ocasiones en que esta palabra ocurre en el corpus. La palabra más frecuente, el pronombre personal de sujeto en primera persona "I" apareció 2.616 veces, mientras que su forma derivada, el pronombre personal de objeto "me", ocurrió 232 veces. La última columna registra que la dispersión de la palabra en el corpus refleja cuantas veces ocurre la palabra en cuántos textos individuales del corpus. Valores de dispersión más bajos representan que la palabra es frecuente en muchos textos, como es el caso de "*I*" con un coeficiente de dispersión de 0,2109 que aparece en todos los textos del corpus desde la frecuencia menor de 228 en un texto hasta 1.288 veces en otro, mientras que valores más altos significan que se distribuye en menos archivos, como la palabra "*you*" con 0,9100, que aparece en 59 de los 68 textos que conforman la muestra, donde su frecuencia en los textos va desde una sola ocurrencia hasta 26 usos.

La mayoría de las palabras en la Tabla 2 son palabras de función como pronombres, preposiciones y conjunciones. Las palabras de contenido más frecuentes son los verbos más comunes como "*like*", "*have*", "*think*" y "*know*", probablemente relacionados con las preguntas en las tareas. Adicionalmente, en las colocaciones más frecuentes aparecen frases aprendidas como "*in order to* X". El participante 9 de la primera muestra de MexLeC usa esta construcción de manera natural en el siguiente ejemplo:

> *I like I am more interested in go to England because one of my aunts go there for I don't know three years to* in order to *teach Spanish*. (MexLeC, UAEMex, participante 9, primer levantamiento)

Se pueden observar datos preliminares de los primeros dos levantamientos de datos en la UAEMex en la Tabla 3:

Tabla 3. Conteos de palabras, tipos y lexemeas para datos MexLeC: UAEMex.

Recolección	Palabras	Tipos	Lexemas
Primero	39.677	2.346	2.270
Segundo	49.218	2.439	2.354

En la tabla se observa un incremento en la producción léxica al comparar las producciones del primero al segundo año de recolección en la universidad, además de un mayor número de unidades utilizadas en cada producción.

4. CONCLUSIONES

Los datos en el corpus MexLeC pueden ser de utilidad para la creación de materiales e intervenciones pedagógicas centradas específicamente en las necesidades de los aprendices universitarios de inglés en general y en el contexto específico. La información brindada por los datos acerca de los índices léxicos de los aprendices, sus errores estructurales, sus recursos y estilos discursivos permiten identificar los apoyos requeridos en sus procesos de aprendizaje, así como las debilidades y fortalezas de los programas educativos. El corte longitudinal de los datos puede ser de gran utilidad en la investigación sobre los procesos de adquisición de rasgos o patrones lingüísticos (fonéticos, fonológicos, morfológicos, sintácticos, semánticos, pragmáticos o discursivos) específicos y la comparación de estos hallazgos con datos de otros estudiantes (nacionales o extranjeros) puede develar la influencia del contexto de aprendizaje, la edad o la lengua materna.

Algunas limitaciones del corpus MexLeC están relacionadas con la falta de textos interactivos (no monológicos) que permitan capturar las estrategias de interacción y los recursos no verbales que se utilizan en la comunicación. Asimismo, la inclusión de tareas con entrenamiento previo permitiría una comparación interesante entre el discurso espontáneo y ensayado. Adicionalmente, para tener una visión más amplia de la lengua producida por aprendices de inglés es necesario incluir discurso escrito en sus diversas variantes y tipos textuales. Una limitación tecnológica para usos con fines de investigación es la falta de una plataforma que permita gestionar y procesar el corpus permitiendo generar listados de frecuencias, concordancias, colocaciones y estadísticos relacionados. Finalmente, es necesario continuar con la recolección de nuevos grupos de estudiantes de diversas regiones del país, esto permitirá la adecuada representación del aprendizaje de inglés en el contexto universitario mexicano.

REFERENCIAS BIBLIOGRÁFICAS

Anthony, Laurence. (2022). TagAnt version 2.0.5. Disponible en: https://www.laurenceanthony.net/software

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. *7es journées internationales d'analyse statistique des données textuelles* (pp. 16-34).

Brezina, Vaclav, Gablasova, Dana y Reichelt, Susan. (2018). BNClab. Disponible en: http://corpora.lancs.ac.uk/bnclab

Brezina, Vaclav, Weill-Tessier, Pierre, y McEnery, Antony. (2021). #LancsBox v. 6.x. [software] Disponible en: http://corpora.lancs.ac.uk/lancsbox/

Butragueño, Pedro Martín y Lastra, Yolanda (coords.). (2012). Corpus sociolingüístico de la Ciudad de México. Vol. II: Hablantes de instrucción media. Ciudad de México: El Colegio de México.

Cambridge University Press. (2023). *Cambridge corpus*. Disponible en: https://www.cambridge.es/nosotros/cambridge-corpus.

Centre for English Corpus Linguistics. (2023). *Learner Corpora around the World*. Louvain-la-Neuve: Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Colantino, Claudia. (2023). El uso de corpus lingüísticos con fines pedagógicos: El caso del SEAH project. *Revista de Lingüística Teórica y Aplicada,* 61(1), 75-92.

CORDE = Real Academia Española: Banco de datos. *Corpus diacrónico del español*. Disponible en https://www.rae.es/banco-de-datos/corde

Council of Europe. (2020). *Common European Framework of Reference for languages: Learning, teaching, assessment. Companion volume with new descriptors.* Strasbourg: Council of Europe.

CREA = Real Academia Española: Banco de datos. *Corpus de Referencia del Español Actual*. Disponible en https://www.rae.es/banco-de-datos/crea

Davies, Mark. (2008). *The Corpus of Contemporary American English (COCA)*. Disponible en: https://www.english-corpora.org/coca/.

Delais-Roussarie, Elisabeth y Yoo, Hi-Yon. (2011). Learner corpora and prosody: from the COREIL corpus to principles on data collection and corpus design. *Poznań Studies in Contemporary Linguistics,* 47(1), 26-39.

Díaz-Negrillo, Ana. (2012). Learner corpora: the case of the NOSE corpus. *Systemics, cybernetics and informatics,* 10(1), 42-47.

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

Ehrnrooth, Lasse. (2015). *FUSE - The Finnish Upper Secondary School Corpus of Spoken English.* Disponible en: https://fusecorpus.eu

Feldweg, Helmut. (1992). *The European Science Foundation Second Language Databank.* Disponible en: https://www.mpi.nl/ISLE/overview/Overview_ESFSLD.html

Flores, Ana y Moore, Pauline. (2023). *Mexican Learner Corpus.* Disponible en: https://sites.google.com/view/mexlec/intro?authuser=0

Gablasova, Dana, Brezina, Vaclav, y McEnery, Tony. (2019). The Trinity Lancaster Corpus: Development, Description and Application. *International Journal of Learner Corpus Research,* 5(2), 126-158.

Gilquin, Gaëtanelle y Meunier, Fanny. (2015). From design to collection of learn-

er corpora en S. Granger, G. Guilquin y F. Meunier (eds.). *The Cambridge handbook of learner corpus research* (pp. 9-34). Reino Unido: Cambridge University Press.

Gilquin, Gaëtenelle, De Cock, Sylvie y Granger, Sylviane. (2010). *Louvain International Database of Spoken English Interlanguage*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Granger, Sylviane. (2002). A Bird's-eye view of learner corpus research. En S. Granger, J. Hung y S. Petch-Tyson (eds.) *Computing, learner corpora, second language acquisition and foreign language teaching* (pp. 3-36). Países Bajos: John Benjamins.

Granger, Sylviane. (2008). Learner corpora en A. Ludeling y M. Kytö (eds.) *Corpus linguistics. An international handbook. Volumen 1* (pp. 259-274). Alemania: Walter de Gruyter.

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

Hasselgren, Angela. S/F. *The EVA Corpus of Norwegian School English*. Disponible en: http://korpus.uib.no/icame/ij21/eva-corp.pdf

Hilton, Heather. (2008). *Parallèle Oral en Langue Étrangère*. Francia: Université de Savoie.

Ishikawa, Shin. (2023). *ICNALE: The International Corpus Network of Asian Learners of English*. Disponible en: http://language.sakura.ne.jp/icnale/

Jucker, Andreas H., Müller, Simone y Smith, Sara. (2006). *GLBCC Giessen - Long Beach Chaplin Corpus*. *Oxford Text Archive.* Disponible en: http://hdl.handle.net/20.500.12024/2506

Kamiyama, Takeki, Lacoste, Véronique y Herry-Benit, Nadine. (2017). The ICE-IPAC project: Testing the protocol on Norwegian and French learners of English. *NewSounds 2016: 8th International Conference.* Disponible en: https://hal.science/hal-01431432

Leech, Geoffrey, Rayson, Paul y Wilson, Andrew. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. Longman.

Lehmberg, Timm y Wörner, Kai. (2008). Annotation stardards. En A. Ludeling y M. Kytö (eds.). *Corpus linguistics. An international handbook. Volumen 1* (pp. 484-500). Alemania: Walter de Gruyter.

Macías, Miguel A. (2020). Análisis de las interferencias léxico-semánticas y morfosintácticas del español al inglés en el texto narrativo de estudiantes universitarios. Tesis Doctoral. Ecuador: Universidad Nacional de Rosario.

Mackey, Alison y Gass, Susan. (2005). *Second language research: Methodology and design*. Lawrence Erlbaum Associates Publishers.

McCarthy, Michael y O'Keeffe, Anne. (2009). Corpora and spoken language en

A. Lüdeling y M. Kytö (eds.). *Corpus linguistics. An international handbook. Volume 2* (pp. 1008-1023). Alemania: Walter de Gruyter.

McEnery, Tony y Gabrielatos, Costas. (2006). English corpus linguistics en B. Aarts y A.

McEnery, Tony y Hardie, Andrew. (2012). *Corpus Linguistics: Method, theory and practice*. Reino Unido: Cambridge.

McEnery, Tony, Xiao, Richard y Tono, Yukio. (2006). *Corpus-based language studies: An advanced resource book*. Reino Unido: Routledge.

McMahon (eds.). *The handbook of English linguistics* (pp. 33-71). Reino Unido: Blackwell.

Meunier, Fanny. (2021). Introduction to learner corpus research. En N. Tracy-Ventura y M. Paquot (eds.). The Routledge handbook of second language acquisition and corpora (pp. 23-36) Reino Unido: Routledge.

Muñoz, Carmen (ed.). (2006). *Age and the Rate of Foreign Language Learning*. Clevedon: Multilingual Matters.

O'Donnell, Michael. (2012). Using learner corpora to redesign university-level EFL grammar education. *Volumen monográfico*, 145-160.

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

Pearson Education. (2022). *The Longman corpus network.* Disponible en: https://www.ldoceonline.com

Salakhyan, Elena. (2012). The Tübingen Corpus of Eastern European English (TCEEE): From a small-scale corpus study to a newly emerging non-native English variety. *Token: A Journal of English Linguistics,* 1, 143-157.

Simpson, Rita C., Briggs, Sarah L., Ovens, J. y Swales, John M. (2002). The Michigan Corpus of Academic Spoken English. Ann Arbor, MI: The Regents of the University of Michigan.

SULEC. (2023). *The Santiago University Learner of English Corpus*. Disponible en: https://sulec.cesga.es/

Tortel, Anne. (2008). ANGLISH. Une base de données comparatives de l'anglais lu, répété et parlé en L1 y L2. *Travaux Interdisciplinaires du Laboratoire Parole et Langage*, 27, 111-122.

UCREL. (2023). *CLAWS part-of-speech tagger for English*. Disponible en: https://ucrel.lancs.ac.uk/claws/

University College London. (2023). *EUROM 1. Multilingual Speech Corpus*. Disponible en: https://www.phon.ucl.ac.uk/shop/eurom1.php

Váradi, Tamás. (2001). The linguistic relevance of corpus linguistics. *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 587-593).

Zapata, Chinger. (2019). Corpus oral de estudiantes de inglés en Chile (ESOC-Chile): diseño, estructura y aplicaciones. *Revista de Lingüística Teórica y Aplicada,* 57(2), 13-40.