RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 57 (2), II Sem. 2019, pp. 13-40.

CL ISSN 0033 - 698X

CORPUS ORAL DE ESTUDIANTES DE INGLÉS EN CHILE (ESOC-CHILE): DISEÑO, ESTRUCTURA Y APLICACIONES[1]

ENGLISH STUDENTS' ORAL CORPUS IN CHILE: DESIGN, STRUCTURE AND APPLICATIONS

CHINGER ZAPATA
Universidad Católica del Norte
czapata@ucn.cl

RESUMEN

El presente artículo constituye el reporte del diseño e implementación del proyecto Corpus Oral de Estudiantes de Inglés en Chile o ESOC-Chile por su sigla en inglés. El ESOC-Chile es el primer corpus de aprendices oficial que registra la producción oral del inglés por parte de estudiantes hablantes nativos del español. Para su construcción, se realizaron grabaciones a través de una entrevista semidirigida, la cual se aplicó a una muestra de 32 informantes de un universo de 180 alumnos de la carrera de Pedagogía en Inglés de la Universidad Católica del Norte en Chile. El resultado muestra una base de datos compuesta por un total de 73631 palabras distribuidas en 3944 tipos de palabras diferentes. El corpus se presenta en tres formatos: textos orales, textos sin etiquetas y textos con etiquetas. Se espera que tanto académicos como estudiantes del inglés puedan utilizar la información contenida en el corpus para realizar investigación sobre la producción oral que conllevará a una mejor comprensión de la competencia comunicativa y, por ende, a una mejor instrucción de la lengua sajona.

Palabras clave: Lingüística de Corpus, corpus de aprendices, inglés como lengua extranjera.

ABSTRACT

This paper reports on the design and construction of a project titled English Students' Oral Corpus in Chile (ESOC-Chile). ESOC-Chile becomes the first learner corpus in the country that holds the EFL oral production of students whose mother tongue is Spanish. Data were collected through interviews. The sample was composed of 32 informants out

[1] El proyecto ESOC-Chile contó con el financiamiento de la Vicerrectoría de Investigación y Desarrollo Tecnológico de la Universidad Católica del Norte, tras la adjudicación del concurso: Proyectos Semilla año 2015.

of a universe of 180 students from the School of English at Universidad Católica del Norte in Chile. The result is a database containing a total of 73631 words distributed in 3944 different types of words in three formats: oral texts, tagged texts and untagged texts. Researchers, teachers and students are expected to study the informants' oral production in the corpus in order to have a better understanding of the students' communicative competence and thus improve the instruction of the language.

Keywords: Corpus Linguistics, learner corpus, English Foreign Language.

Recibido: 17/10/2018. Aceptado: 28/11/2019.

1. INTRODUCCIÓN

La utilización de un corpus es, hoy por hoy, uno de los recursos más versátiles en el estudio y la enseñanza de cualquier lengua nativa o extranjera (Römer, 2009). A través de un corpus se puede estudiar aspectos fonético-fonológicos, morfosintácticos, léxico-semánticos, léxico-gramáticos y pragmático-discursivos del uso de las lenguas, desde una perspectiva tanto diacrónica como sincrónica (Bentivoglio y Malaver, 2006). Por esta razón, muchas organizaciones y universidades a nivel nacional e internacional dedicadas a los estudios de lingüística pura y aplicada han invertido recursos en la construcción de sus propios corpora. Entre los más representativos de la lengua inglesa tenemos el COCA[2] (2008-) de la Universidad Brigham Young, y el OANC[3] (1990-2016) de la organización *American National Corpus Project*, ambos en los Estados Unidos; por su parte, en Inglaterra tenemos el COBUILD[4] (1980) de la Universidad de Birmingham y el BNC[5] (2007) de la Universidad de Oxford. En Chile existe hasta la fecha la propuesta de creación de un corpus chileno de inglés hablado como idioma extranjero de Ortega (2014), donde se prevé estudiar los errores típicos de los estudiantes chilenos de pedagogía en inglés; pero desde la publicación de la primera etapa de su trabajo no se ha tenido más noticias al respecto, por lo que se presume aún en construcción. Este vacío nos ha llevado a construir un corpus oral de aprendices del inglés como lengua extranjera.

Pero ¿qué es un corpus y por qué es importante que se construya uno? Para responder estas interrogantes se recurrió en primer lugar a Stubbs (2002), quien señala:

[2] COCA: Corpus of Contemporary American English
[3] OANC: Open American National Corpus
[4] COBUILD: Collins Birmingham University International Language Database.
[5] BNC: British National Corpus

> Un corpus es una colección de textos diseñado con algún propósito, por lo general, para la enseñanza o la investigación. Un corpus no es algo que el hablante haga o sepa, sino algo que construye un investigador. Es entonces un registro de la actuación, por lo general de muchos usuarios, concebido para su estudio y poder así hacer inferencias acerca del uso típico del lenguaje [Mi traducción = MT] (p. 239)[6].

Con respecto a la importancia de la construcción de un corpus, McEnery y Wilson (2001) sostienen que esta colección de muestras del lenguaje cotidiano en uso (datos empíricos) representa una de las fortalezas de los corpora, ya que hacen el análisis lingüístico más objetivo.

Por otro lado, el impacto positivo de la utilización de los corpora en la enseñanza de lenguas maternas y extranjeras ha sido ampliamente comprobado. Los aportes en la enseñanza de lenguas extranjeras pueden clasificarse de dos maneras: aplicaciones indirectas y aplicaciones directas (Leech, 1997; Torruella y Llisterri, 1999; Hunston, 2002; Nesselhauf, 2004; Bennett, 2010; McEnery y Xiao, 2010; Reppen, 2010).

Las aplicaciones indirectas pueden estar relacionadas, por un lado, al desarrollo y diseño de materiales instruccionales como guías de vocabulario, colocaciones, guías de estudio, diseño de estrategias didácticas a partir de los errores típicos encontrados, entre otros. Por el otro, a decisiones acerca del contenido de los programas de estudio con base en muestras de lenguaje (oral o escrito) provenientes de corpus y sobre la base de las evidencias lingüísticas.

En relación con las aplicaciones directas, se deben considerar dos actores fundamentales y un recurso: docentes, alumnos y el programa de concordancias. Tanto docentes como alumnos pueden utilizar los datos del corpus como parte de la estrategia de enseñanza y aprendizaje de algún aspecto particular de la lengua.

En este escenario, los docentes adoptan un papel de tutor donde se instruye al estudiante en un proceso de enseñanza basado en la investigación. En otras palabras, el docente, apoyado en un programa de concordancias, crea el escenario para que los estudiantes desarrollen estrategias de aprendizaje que facilitan comprender aspectos de la lengua, a través del acceso directo a datos lingüísticos.

El programa de concordancias constituye, por ende, un recurso computarizado al interior de la sala de clases, que procesa muestras de lengua oral o escrita producidas por los mismos estudiantes. Este tipo de programa permite observar patrones de uso de la lengua de tipo sintáctico, establecer concordancias para determinar colocaciones léxicas, calcular frecuencias de uso de ciertos patrones o lexías, entre otras acciones.

[6] Original en inglés: A corpus is a collection of texts, designed for some purpose, usually teaching or research. […] A corpus is not something that a speaker does or knows, but something constructed by a researcher. It is a record of performance, usually of many different users, and designed to be studied, so that we can make inferences about typical language use.

Por su parte, este procedimiento permite a los estudiantes tener un grado mayor de autonomía en relación a su proceso de instrucción ya que aprenden a manejar un recurso que les permite analizar la producción oral o escrita de los hablantes; es decir, observan patrones sintácticos, estructuras lingüísticas repetitivas, las colocaciones más comunes para un determinado vocablo empleados por ellos mismos en el contexto oral o escrito, errores típicos de sintaxis oral, de pronunciación, de ortografía, análisis del ritmo de las oraciones, entre otros, los cuales podrían compararse con los usos de los hablantes nativos en corpora como el COCA o el COBUILD. A este tipo de práctica, Johns (1991) lo ha denominado *aprendizaje a partir de los datos* o DDL por sus siglas en inglés: *data-driven learning*. De acuerdo con el autor, este enfoque considera que: "…el aprendiz de la lengua también es, en esencia, un investigador cuyo aprendizaje debe ser generado a través del acceso a los datos lingüísticos" [Mi traducción = MT] (p. 2)[7].

En cualquiera de los dos casos, bien sea desde su aplicación directa o indirecta, no cabe duda de que la construcción de un corpus constituye una fuente interminable de datos para el estudio de la competencia comunicativa de los estudiantes, competencia esta entendida como el conocimiento subyacente que cada hablante en particular posee sobre la estructura formal y los patrones de uso de la lengua (Hymes, 1992).

La incorporación de un corpus en el ámbito de la enseñanza de lenguas impulsaría por efecto el trabajo de investigación en distintas áreas como la fonética y la fonología, la morfología, la sintaxis, el vocabulario, la pragmática, la semántica, los cuales darían pie para concebir nuevas estrategias didácticas, el desarrollo de materiales de enseñanza e información empírica y teórica de referencia para el diseño o rediseño curricular, entre otros. Sobre las aplicaciones directas, se hará referencia más amplia en la sección 4 de este artículo.

Ahora bien, la mayoría de los corpora existentes en cualquier lengua ha sido diseñado sobre la base de hablantes nativos. Si bien es cierto que sus aportes, de manera directa e indirecta, en la enseñanza y aprendizaje de lenguas extranjeras han sido de gran ayuda, también es cierto que es necesario tener una base de datos que arroje luz en cuanto a la producción oral de estudiantes de inglés como lengua extranjera, cuya lengua madre sea el español.

Como ya se había acotado, no existe en Chile ni en ningún otro país (por lo menos no existen registros oficiales hasta la presente fecha) un corpus actualmente operativo y público de esta naturaleza. Es por ello que el presente artículo tiene como objetivo principal reportar sobre el diseño, desarrollo, estructura y aplicaciones del proyecto: Corpus Oral de Estudiantes de Inglés en Chile (*English Students' Oral Corpus in Chile,* en adelante ESOC-Chile). Esta iniciativa llevó a cabo

[7] Original en inglés: The language-learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data.

la construcción de un banco de datos constituido exclusivamente por hablantes no nativos del inglés cuya lengua materna es el español. Se espera que este corpus brinde tanto a académicos como estudiantes la oportunidad de analizar la producción oral de la lengua extranjera de estudiantes de la carrera de Pedagogía en Inglés de la Universidad Católica del Norte (UCN) en Chile. El propósito es tener una mejor comprensión del aprendizaje, desarrollo y producción de las competencias comunicativas de los estudiantes, que permitan ajustar o mejorar la instrucción y/o el diseño de contenidos de cursos y materiales didácticos que estén en consonancia con las necesidades particulares contextualizadas de los aprendices de inglés. De este modo, el ESOC-Chile pasa a conformar la primera base de datos genuinos de producción oral del inglés de manera pública y oficial en la región y en el país.

2. DISEÑO Y ESTRUCTURA DEL ESOC-CHILE

Para llevar a cabo la construcción del corpus, se siguieron los siguientes pasos:

1. Recolección de las muestras orales a partir de grabaciones digitales por medio de entrevistas personales.
2. Almacenamiento de las grabaciones en formato de audio digitalizado para su posterior procesamiento.
3. Transcripción ortográfica de las grabaciones obtenidas para conformar el banco de datos.
4. Codificación de las transcripciones para el registro, control y manejo de los datos.
5. Etiquetaje de las transcripciones para su procesamiento con herramientas computarizadas (programas de concordancia).

2.1. Metodología

La metodología para la construcción del corpus se dividió en dos etapas. La primera fue la etapa de recolección de datos; la segunda, el procesamiento de los datos.

2.1.1. Primera etapa

Esta etapa estuvo orientada a la selección de los informantes y los criterios que se emplearon para la recolección de los datos.

- *Selección de los informantes*

Para la selección de los informantes se aplicó la afijación uniforme; es decir, se segmentó el universo en cuotas de acuerdo con las variables sociales que se tomaron en cuenta y se asignó a cada cuota un número igual de informantes (Bentivoglio y Malaver, 2012). El universo estuvo compuesto por 180 estudiantes universitarios de la carrera de Pedagogía en Inglés de la UCN. No obstante, se trabajó con una muestra del universo de 32 estudiantes (17% del total de estudiantes), cuya distribución fue equitativa: ocho informantes por año.

Por otro lado, la selección de los informantes también se hizo siguiendo los siguientes criterios. Los informantes: a) deben ser chilenos de nacimiento, así como sus padres y sus abuelos, b) deben ser estudiantes universitarios de una carrera relacionada al estudio del inglés, en este caso: Pedagogía en Inglés, c) nunca deben haber viajado a ningún país de habla inglesa, d) no deben tener la influencia del idioma extranjero fuera del contexto educacional, por ejemplo familiares directos angloparlantes, e) deben tener todas sus piezas dentales completas, y e) no pueden presentar ningún trastorno de producción oral como el Síndrome de Tourette o trastornos de comunicación como la tartamudez u otro parecido.

- *Las variables sociales*

En el diseño de un corpus donde se analice el habla de los informantes es necesario tomar en cuenta las variables sociales para su selección, por cuanto las mismas responden a "…la necesidad de que el corpus refleje las características sociológicas generales de la comunidad de habla" (Bentivoglio y Malaver, 2012, p. 153). Considerando que el ESOC-Chile busca analizar la producción oral de hispanohablantes, estudiantes de inglés como lengua extranjera, se cree que la incorporación de ciertas variables sociales podría aportar información de interés para algunos investigadores. Por esta razón, al levantar la base de datos se ha tomado en cuenta dos variables sociales que pudieron ser fácilmente constatadas en el corpus: 1) grado de instrucción y 2) sexo.

El grado de instrucción corresponde al año que cursan los informantes dentro de la carrera de Pedagogía en Inglés. Esta variable está asociada a los niveles de competencia comunicativa que se encuentran alineados con el Marco Común Europeo de Referencia para las Lenguas (Council of Europe, 2014), que están en consonancia con el año en curso de los estudiantes de acuerdo a los resultados de aprendizaje descritos en los programas de curso de las asignaturas Discurso Oral y Escrito en Inglés. En este sentido, se presume que los estudiantes de primer año están en su proceso de desarrollo de un nivel B1; los de segundo y tercer año en un proceso de desarrollo del nivel B2 y los de cuarto año en desarrollo del nivel C1.

La segunda variable, el sexo, se tomó en consideración por cuanto se puede

hacer una distinción de las diferencias significativas que existen en la producción de la lengua entre hombres y mujeres (Trudgill, 1972; Lakoff, 1975; Zimmerman y West, 1975; Coates, 1986; Eckert, 1989; Tannen, 1990; Labov, 1994, 2001; Cameron, 1995; Talbot, 1998; Holmes y Meyerhoff, 1999; Eckert y McConnell-Ginet 2003; Thi Ngoc, 2013; Ehrlich, Meyerhoff y Holmes, 2014). De este modo, los ocho alumnos por cada año se distribuyen en cuatro hombres y cuatro mujeres. El total de informantes quedó distribuido como se muestra en la Tabla I.

Tabla I. Distribución de los informantes por nivel de competencia comunicativa y sexo.

Años / **Sexo**	1er Año (B1)	2do Año (B2)	3er Año (B2)	4to Año (C1)	**Total**
Mujer	4	4	4	4	**16**
Hombre	4	4	4	4	**16**
Total	**8**	**8**	**8**	**8**	**32**

- *El instrumento*

Para recoger los datos se llevó a cabo una entrevista semiestructurada o, en términos lingüísticos, el método de conversación semidirigida (Silva, 2001), las cuales consistieron en grabaciones de 15 minutos. El propósito de la misma fue permitirle al informante hablar continuamente para que produjera un registro lo más espontáneo y natural posible.

- *El procedimiento*

En cada entrevista participaron un entrevistador (E), una persona entrevistada la cual toma el papel de informante (I) y una persona como audiencia (A) que solo se limitó a observar la conversación y llevar un registro silencioso del tiempo y del proceso de la entrevista. Durante las entrevistas el entrevistador realizó una serie de preguntas abiertas (ver Anexo 1) para incitar a la conversación, evitando interrumpir en la medida de lo posible la participación del informante. Las preguntas versaban sobre tópicos de interés de los informantes.

- *Recolección de datos*

Para la recolección de datos, se contó con la participación de dos hablantes nativos. Los hablantes nativos cumplieron el rol de entrevistadores, facilitando una

conversación fluida y espontánea de manera que las muestras sean auténticas. Cada entrevista tuvo una duración de 15 minutos por persona. La entrevista se realizó en un lugar cerrado para evitar, en la medida de lo posible, el ruido de fondo, pero que propiciara un ambiente de comodidad. El equipo que se utilizó fue una grabadora digital en formato MP3, marca Panasonic, modelo RR-US571, la cual aseguró que la grabación tuviera buena calidad para que los datos recogidos se escucharan de forma clara y sin ningún inconveniente. Este equipo de memoria extendida tiene funciones que le permiten almacenar y transferir los archivos de audio MP3, a través de un puerto USB a computadoras personales.

2.1.2. Segunda etapa

Esta segunda etapa muestra información sobre el procedimiento y procesamiento de los datos.

- *El equipo (recurso humano)*

El equipo del ESOC-Chile está conformado principalmente por Chinger Zapata, académico de la Escuela de Inglés de la UCN y director del proyecto, así como las hablantes nativas del inglés: María Cecilia Ávila y Emily Noble, ambas también académicas de la misma Escuela, quienes fungieron como colaboradoras en el proyecto y llevaron a cabo el levantamiento de la base de datos a través de entrevistas orales. El proceso de transcripción, codificación y etiquetado estuvo a cargo de un grupo de estudiantes de tesis del cuarto año de la carrera de Pedagogía en Inglés que apoyaron el proyecto: Rafael Arríquez Bravo, Claudia Robles Seura, Simone Larrondo Veloso y Milca Llantén Toledo. Estos procesos también contaron con la supervisión y revisión del investigador principal del ESOC-Chile.

- *La codificación*

La codificación de los archivos siguió las siguientes pautas. Se utilizó un sistema que comienza con las siglas identificativas de la comunidad estudiada (ejemplo: EIUCN: Escuela de Inglés – Universidad Católica del Norte), seguidos de un número de dos cifras que indica el orden en que fueron entrevistados los informantes, comprendido entre el 01 y el 32, número máximo de informantes utilizados, luego el código sociolingüístico del informante que indica la variable sexo M (mujer), H (hombre) y posteriormente el código de la variable grado de instrucción: 1A (1er año), 2A (2do año), 3A (3er año), 4A (4to año), quedando de la siguiente manera: EIUCN_01_M1A. Esta codificación es válida para los propósitos del corpus mismo; no obstante, cada equipo de investigación que utilice el corpus podrá adjudicar a los archivos otros códigos y numeraciones para sus fines particulares.

- *La transcripción*

La transcripción de los datos recogidos durante las entrevistas se realizó en formato Word. Es importante destacar que las transcripciones se realizaron de forma literal, es decir que cada error, sonido o vacilación que el entrevistado expresó se transcribió. Esto último por la razón de que cualquiera de estas acciones puede significar un dato importante para futuras investigaciones.

- *El etiquetado*

Una vez que los datos fueron codificados y transcritos, se procedió a etiquetarlos. Este proceso se realizó por medio del método *text encoding initiative,* TEI. De acuerdo con el TEI Consortium (2016), el TEI:

> …es un consorcio el cual desarrolla y mantiene de manera colectiva un estándar para la representación de textos en forma digital. Su servicio principal lo constituye una serie de lineamientos a través de los cuales se especifican métodos de codificación para textos que serán leídos por programas de concordancia, empleados principalmente en las ciencias sociales, en las humanidades y en la lingüística [MT] (s/n)[8].

- *Configuración de los textos*

Los textos se presentan en tres formatos: texto en audio, texto con etiquetas, texto sin etiquetas. Los textos etiquetados presentan dos partes bien diferenciadas: la cabecera y el texto propiamente dicho. Para la elaboración de la cabecera se utilizó una plantilla que contiene datos específicos.

- *La cabecera*

Esta sección incluye información relacionada a los siguientes aspectos:

- Datos del propio archivo
- Datos de la grabación de la entrevista
- Datos sobre la transcripción y revisión de la entrevista
- Datos sobre los hablantes participantes en la entrevista.

[8] Original en inglés: *…is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics.*

Esos datos tienen un formato común que permite a cualquier investigador llevar a cabo un tratamiento homogéneo. Además, la plantilla tiene un formato compatible con XML, para asegurar la recuperación de la información. Los datos específicos se colocan entre comillas en los espacios destinados para tales fines (ver Anexo 2).

- *Marcas y etiquetas del texto*

Los textos de las entrevistas fueron transcritos en ortografía convencional. No obstante, el texto transcrito también incorpora una serie de signos que no son habituales en la escritura ordinaria. Los mismos indican aspectos puntuales de la representación escrita de la lengua oral. El marcado y etiquetado del texto utilizado para el ESOC-Chile ha sido el mismo método utilizado por el corpus PRESEEA (2014), que a su vez se apoyó en el método TEI. Este método, de acuerdo con Bentivoglio y Malaver (2012), consiste en un grupo de marcas o etiquetas que se asignan al discurso tanto de informantes como de entrevistadores. Las mismas incluyen símbolos para identificar el texto, así como también símbolos para señalar otros elementos del texto tales como exclamaciones, interrogaciones, sonidos onomatopéyicos, nombres propios, citas directas, énfasis, alargamientos, silencios, extranjerismos, siglas, risas, vacilaciones e incluso ruidos producidos por los participantes o ruidos del entorno. Estas etiquetas enmarcan todo el texto con los siguientes símbolos de apertura y cierre (< / >). Las marcas y etiquetas comunes del ESOC-Chile se presentan en el Anexo 3.

Con respecto a los textos sin etiquetas, su finalidad es la publicación impresa de los materiales y su lectura convencional. La presentación de un texto sin etiquetar consiste básicamente en disponer de una cabecera con los datos esenciales del informante, el número de entrevista, la fecha de la grabación y el texto de la transcripción desprovisto de las etiquetas, tanto las de apertura y cierre, como las aisladas, excepto <risas = " "/> y <silencio/> (ver Anexo 4).

- *Revisión*

Los procesos de codificación, transcripción y etiquetado contaron con la supervisión del investigador principal del ESOC-Chile mencionado en la sección *2.1.3.1*. A la presente fecha, a los datos se les ha realizado tres revisiones generales para descartar errores de transcripción y etiquetado. Los datos han quedado registrados en tres formatos: 1) texto etiquetado (para su manejo a través de medios informáticos con programas de concordancias como el AntConc), 2) texto sin etiquetas (para la lectura de aspectos generales), y 3) textos en audio (para el análisis personalizado de cada investigador).

3. EL CORPUS DISEÑADO

Después de la compilación y procesamiento de los datos, el corpus ha quedado confeccionado. A continuación se presenta sus características. El ESOC-Chile posee una base de datos oral, compuesta por 73631 palabras (tokens)[9] y 3944 tipos diferentes de palabras (*types*)[10] en un registro de habla espontánea e informal. Tiene un total de 32 informantes distribuidos de acuerdo a dos variables sociales: La primera, el grado de instrucción y la segunda: el sexo, la cual distribuye a los informantes de manera equitativa en 16 hombres y 16 mujeres.

3.1. El tipo de corpus

De acuerdo con sus especificaciones y por el uso que se puede hacer de él, el ESOC-Chile es un corpus de aprendices. Nesselhauf (2004) define el corpus de aprendices como: "…colecciones de textos computarizados y sistemátizados producidos por aprendices de una lengua…" [MT] (p. 125)[11]. Básicamente, el ESOC-Chile constituye un conjunto de textos orales en inglés producidos por hablantes nativos del español, los cuales han sido procesados y almacenados digitalmente para su estudio. En una perspectiva más amplia, Baker et. al. (2006) no solo definen, sino que también describen su utilidad, al señalar que:

> los corpus de aprendices son útiles en estudios de adquisición de una segunda lengua ya que ayudan a construir el perfil lingüístico de los aprendices, específicamente en relación a análisis de errores o para indagar sobre qué palabras, frases, categorías gramaticales, entre otras, son empleadas con mayor o menor frecuencia por los aprendices en comparación con los hablantes nativos [MT] (p. 103)[12].

[9] Según Baker, Hardie y McEnery (2006), la palabra token se define como: "Una unidad lingüística, a menudo una palabra…" (p. 161). Original en inglés: *A single linguistic unit, most often a word…*

[10] Types se define como: "…el número de tipos se refiere al número total de palabras únicas. Por ejemplo, la palabra *barco* puede ocurrir 177 veces en un corpus, pero solo se cuenta como un tipo de palabra" (p. 162). Original en inglés: *"…the number of types refers to the total number of unique words. For example, the word ship may occur 177 times in a corpus, but it only counts as one type of word."* (Baker et al., 2006)

[11] Original en inglés: *…systematic computerized colections of texts produced by language learners…*

[12] Original en inglés: *A learner corpus consists of language output produced by learners of a language. […] Learner corpora are useful in studies of second language acquisition as they help to build a profile of learner language, particularly in terms of error analysis or for ascertaining what words, phrases, parts-of-speech etc. are over- or under-used by learners, compared to native speakers.*

En este mismo orden de ideas, el Instituto Cervantes (2016) señala que a través de este tipo de corpus se puede percibir los niveles de competencia comunicativa de los aprendices con respecto a la lengua que aprenden, lo cual es uno de los propósitos fundamentales para los cuales se ha construido este corpus.

Por otro lado, considerando los criterios de clasificación de Torruella y Llisterri (1999), el ESOC-Chile puede ser, en primer lugar, un *corpus monitor,* clasificación ésta que se hace a partir del porcentaje y distribución de los textos. Un *corpus monitor* es aquel que mantiene una cantidad constante de volumen textual que se actualiza cada cierto tiempo. Por tanto, en la medida que se van incorporando nuevos textos al cabo de un período de tiempo, también se van desincorporando otros. El ESOC-Chile tiene ese propósito. De este modo, los textos que se desincorporan pasarán a formar una base de datos con los que posteriormente se puede construir otro corpus de tipo *diacrónico* y poder realizar estudios sobre las competencias comunicativas de los informantes en distintas generaciones de estudiantes.

En segundo lugar, el ESOC-Chile se clasifica también como *corpus documentado,* la cual tiene que ver con la documentación que acompaña a los textos en la cabecera. Finalmente, la última clasificación se relaciona con los criterios específicos para la clasificación de corpora orales. En este sentido, el ESOC-Chile se considera un corpus de tipo *transcripciones ortográficas de la lengua hablada.* Torruella y Llisterri (1999) lo explican de la siguiente manera:

> En la lingüística de corpus tradicional se ha trabajado habitualmente con transcripciones ortográficas de la lengua hablada, procedentes de entrevistas realizadas especialmente para el corpus, de conversaciones espontáneas o de los medios de comunicación, incluyéndose también otros materiales propios del registro oral como discursos políticos, clases, sermones, etc. Aunque el punto de partida sea una grabación, una vez transcrito, el corpus se trata con los mismos procedimientos que un corpus textual… (p. 15).

3.2. Limitaciones del corpus

Es importante señalar que, por ser esta una primera edición del ESOC-Chile cuya vigencia va de 2015 a 2018 (duración de la carrera de los estudiantes del primer año del corpus), existen dos limitaciones: la primera, el tamaño de la muestra (32 informantes) que incluye el número total de palabras y tipos de palabras generadas; la segunda, la certeza del nivel de competencia comunicativa de los informantes. Con respecto a la muestra, esta selección pequeña constituye una limitación, ya que la misma no es representativa del universo para establecer generalizaciones con respecto a los errores típicos y comunes de los estudiantes de acuerdo con el nivel de competencia comunicativa.

En relación al nivel de competencia comunicativa de los informantes, para esta primera edición, se ha asumido que los estudiantes que cursan cada año poseen el nivel de competencia comunicativa que corresponde a los establecidos en los programas de curso de las asignaturas; es decir, B1 para Discurso Oral y Escrito en Inglés (primer año), B2 para Discurso Oral y Escrito en Inglés (segundo y tercer año) y C1 para Discurso Oral y Escrito en Inglés (cuarto año). No obstante, no hay certeza de que el estudiante que cursa cualquiera de estos años posee en efecto el nivel de competencia comunicativa correspondiente.

A pesar de estas limitaciones, la primera edición del corpus ha sido un ejercicio que ha permitido establecer las bases para la construcción de un corpus y la incorporación de la cultura de lingüística de corpus entre los participantes del proyecto, los académicos de la unidad y los estudiantes en general. De hecho, esta primera edición ya ha generado las primeras investigaciones sobre el uso del inglés por parte de estudiantes hablantes nativos del español.

Para superar estas limitaciones, actualmente se está organizando la segunda edición que no solo incluye la producción oral, sino que también incluye la producción escrita. La misma se estima comience en el segundo semestre de 2019 y tendrá una vigencia de 5 años; es decir, de 2019 a 2023. Con un universo total de 238 estudiantes para el primer semestre de 2019, esta segunda edición incluirá como muestra 200 estudiantes aproximadamente (50 para primer año, 60 para segundo año, 40 para tercer año y 50 para cuarto año), lo que representa un 84% del universo. De este modo, garantizamos la representatividad del corpus que nos permita realizar generalizaciones en los hallazgos.

Por otro lado, se aplicará un mock test de la Universidad de Cambridge (PET para primer año, FCE para segundo y tercer año, CAE para cuarto año) para asegurar que los estudiantes que finalmente conformen la muestra sean informantes que en efecto poseen el nivel de competencia comunicativa correspondiente al año que cursan. La aplicación del mock test está planificado para el final del primer semestre de 2019. De este modo, los resultados relacionados a errores típicos y comunes pueden correlacionarse con niveles de competencia comunicativa.

Los datos del ESOC-Chile se actualizarán constantemente cada cinco años. Por otro lado, los datos que se vayan desincorporando pasarán a formar parte de un corpus diacrónico que recogerá la producción del inglés por cortes de estudio de los informantes, proyecto en el cual se trabaja simultáneamente con la producción de la segunda edición del ESOC-Chile.

4. INVESTIGACIONES Y APLICACIONES

El ESOC-Chile es una fuente de información que puede ser utilizada con dos propósitos. El primero se relaciona con la investigación sobre la producción oral

de los estudiantes de la Escuela de Inglés. El segundo, con aplicaciones directas como recurso de enseñanza al interior de la sala de clases.

4.1. Estudios realizados

Dentro de los estudios que se pueden realizar con el corpus están: estudios de transferencia lingüística, estudios de corte descriptivo, estudios contrastivos, entre otros.

A la fecha se han realizado cuatro estudios sobre el discurso oral de los estudiantes de inglés por parte de alumnos que trabajan en el desarrollo de sus tesis de pregrado. Tres de ellas versan sobre frases preposicionales y una sobre el grupo nominal. La primera investigación se tituló: *Uso de las frases preposicionales en el discurso oral de estudiantes de inglés* y se completó en diciembre de 2016. Su objetivo principal fue: *analizar el uso de las preposiciones en el discurso oral de los estudiantes de primer año en la carrera Pedagogía en Inglés de la UCN*. La segunda investigación llevó por título: *Errores comunes en el uso de las frases preposiciones en inglés* y fue realizada en 2017. En ella se propuso como objetivo: *identificar errores en el uso de las preposiciones presentes en el discurso oral de estudiantes de inglés*. La tercera fue un estudio documental de las frases preposicionales, el cual se tituló: *Linguistic Awareness of the Prepositional Phrase Complexities in the EFL Contex*t. Su objetivo principal: *to raise language awareness of the multifaceted nature of the prepositional phrase among teachers and students.* Esta investigación fue realizada por uno de los académicos de la Escuela a partir de los referentes teóricos empleados en los trabajos previos. Aunque no se trabaja en ella con los datos del corpus, la misma surge a partir de las investigaciones anteriores en un proyecto general sobre el tópico en investigación. La cuarta terminó en diciembre de 2018 y se centró en describir estructuras y funciones sintácticas empleadas por los estudiantes en el uso del grupo nominal.

Las tres primeras han sido enviadas a revistas para su publicación y la cuarta está en proceso de revisión para su envío. Otros temas que se abordarán a futuro son:

- Descripción del grupo adjetival
- Descripción del grupo adverbial
- Descripción del grupo verbal
- Estudio contrastivo de la entonación de enunciados declarativos en inglés-español
- Errores típicos de pronunciación
- Inventario léxico de la producción oral de los estudiantes de inglés
- Marcas léxicas de posicionamiento en el discurso oral de los estudiantes de inglés

- Los marcadores del discurso en la producción oral de estudiantes de inglés.

4.2. Aplicaciones directas

El segundo propósito es la utilización de la fuente de datos como recurso didáctico para la promoción del aprendizaje autónomo. Desde esta perspectiva, el enfrentar a los estudiantes con los errores típicos hallados en el corpus les permitirá tomar conciencia sobre qué aspectos deben revisar en su propia producción para mejorarla y evitar los errores comunes.

4.2.1. Consideraciones previas a las aplicaciones directas

- *Programa de concordancias*

Para el desarrollo de actividades con aplicaciones directas al interior de la sala de clases es necesario instruir a los estudiantes en el uso de un programa de concordancias.

De acuerdo con Bennett (2010): "Los programas de concordancias son softwares computarizados usados para acceder y ordenar el corpus" [MT] (p. 16)[13]. A través de ellos se puede procesar la información contenida en el corpus y realizar distintas acciones para obtener frecuencias de palabras, rangos, listas de palabras, cantidad total de palabras, tipos de palabras, palabras en contexto, colocaciones, palabras claves, etc. En la actualidad existen muchos programas de concordancias tales como el MICASE, el WordSmith Tools, el TextStat, el MonoConc; no obstante, la mayoría de estos programas tienen un costo. Hoy en día, uno de los más versátiles y fáciles de usar es el AntConc diseñado por Anthony (2019). Además de cumplir con todas las funciones típicas de los programas de concordancias antes nombrados, el AntConc es libre de costos y se encuentra de forma gratuita en la Internet.

- *El AntConc*

El AntConc, como programa de concordancias, es una plataforma de usos múltiples para desarrollar investigación en lingüística de corpus y realizar actividades para el aprendizaje de una lengua a partir de los datos. Posee las siguientes siete herramientas:

[13] Original en inglés: *Concordancing programs are computer software used to access and sort data from the corpus.*

- **Concordancias (Concordance Tool):** Muestra resultados de búsqueda en el formato "palabras claves en contexto" o KWIC por sus siglas en inglés. Le permite al investigador ver las palabras y frases comúnmente usadas en el corpus.
- **Concordancias en barra (Concordance Plot):** Muestra los resultados de búsqueda trazados como "códigos de barra", permitiendo identificar la posición o lugar donde se ubican los resultados en el texto.
- **Vista del archivo (File View):** Muestra el texto de archivos individuales; de este modo, es posible indagar con mayor profundidad los resultados generados a través de otras herramientas del AntConc.
- **Agrupaciones/N-Grams (Clusters/N-Grams):** La herramienta "agrupaciones" muestra conjuntos de palabras bajo ciertas especificaciones y/o condiciones. En esencia, resume los resultados generados en otras herramientas como "concordancias" o "concordancias en barra". La herramienta N-Grams, por su parte, realiza un escaneo del corpus completo para mostrar la longitud de los conjuntos de palabras (por ejemplo, una palabra, dos palabras…), lo cual permite al investigador conseguir expresiones comunes en el corpus.
- **Colocaciones (Collocates):** Muestra las colocaciones de un término en la búsqueda con lo cual es posible investigar patrones no secuenciales en la lengua.
- **Lista de palabras (Word List):** Esta herramienta cuenta todas las palabras del corpus y las presenta en una lista ordenada, permitiendo encontrar rápidamente cuáles son las palabras más frecuentes en el corpus.
- **Lista de palabras clave (Keyword List):** Muestra cuáles palabras son o no frecuentes en el corpus en comparación con las palabras de un corpus de referencia, con lo que es posible identificar palabras características en el corpus como por ejemplo las que son parte de un género discursivo específico.

4.2.2. Actividades sobre aplicaciones directas

Una vez consolidada la etapa de instrucción del programa de concordancias, se pueden realizar una serie de actividades orientadas a promocionar el aprendizaje autónomo de los alumnos. Al respecto, existe una gran cantidad de materiales, recursos y libros de textos que muestran formas diversas en que los profesores y alumnos pueden utilizar la información de un corpus de aprendices como recurso didáctico para la enseñanza de lenguas (*Corpus Linguistics for English Teachers: Tools, Online Resources, and Classroom Activities* (Friginal, 2018), *Has Corpus-Based Instruction Reached a Tipping Point? Practical Applications and Pointers for Teachers* (Huang, 2017), *Application of Learner Corpora to Second Language Learning and Teaching: An Overview* (Xu, 2016), La Lingüística de Corpus y su incidencia en la enseñanza de lenguas extranjeras (Zapata, 2015), entre otros). En esta sección se sugiere una de esas actividades a modo de ejemplo.

- ***Actividad 1: Corrigiendo la conjugación verbal***

Recurso: ESOC-Chile
Programa de Concordancias: AntConc
Tópico: Conjugaciones verbales
Estudiantes a los que va dirigida la actividad: estudiantes de cualquier año.

Procedimiento:

- Seleccione un verbo cuya frecuencia sea alta en el corpus.

Solicite a los estudiantes seleccionar un verbo de alta frecuencia; es decir, que esté ubicado entre las primeras treinta palabras de la lista (Figura 1). Una vez seleccionado el verbo, pídale a los estudiantes que hagan click sobre el verbo para ver las líneas de concordancias (Figura 2). Para los efectos de este ejercicio se ha seleccionado al verbo *was* como ejemplo.

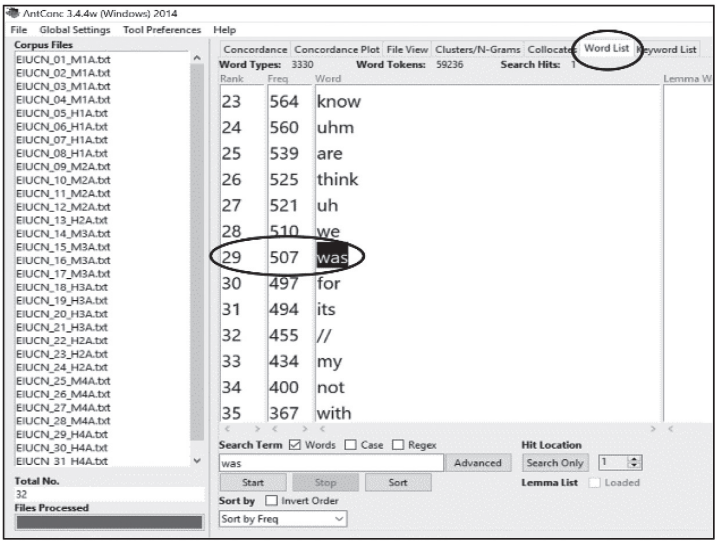

Figura 1. Actividad 1: Corrigiendo la conjugación verbal.

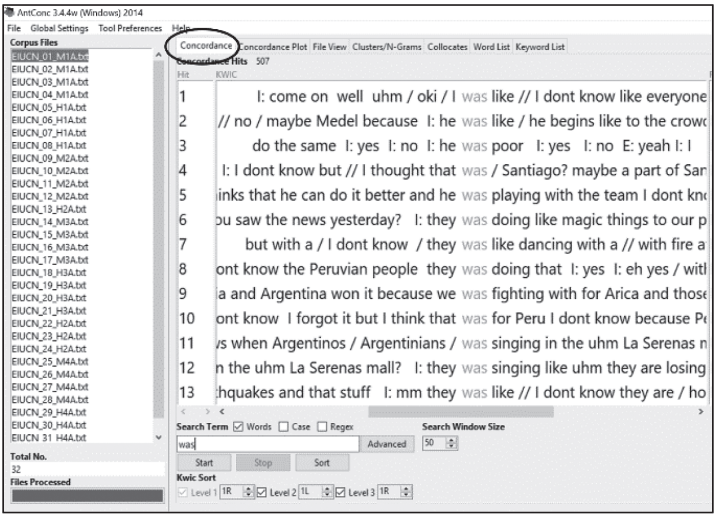

Figura 2. Ejemplo de concordancia.

- Utilice la herramienta ***vista del archivo (File View).***

Una vez que tenga las líneas de concordancias, indique a los estudiantes hacer click sobre el primer verbo de la lista para verlo como texto en el archivo, tal y como se muestra en la Figura 3. Esta opción le permitirá al estudiante ir viendo caso a caso en el contexto de las oraciones completas del archivo donde aparece.

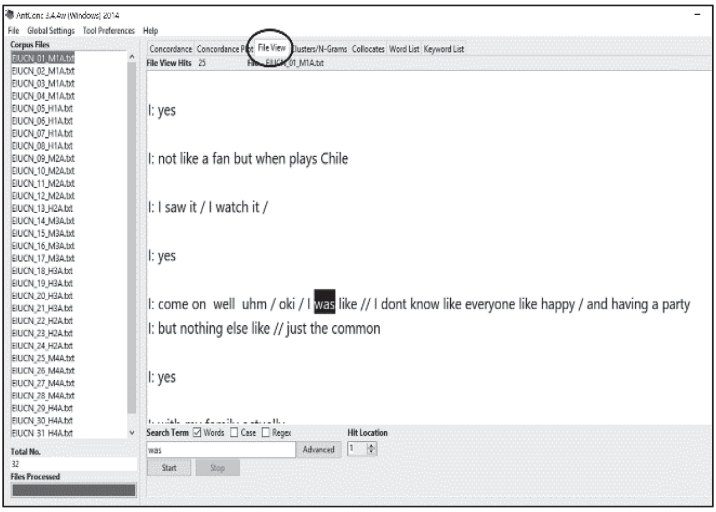

Figura 3. Ejemplo modo texto en el archivo.

A través de la tecla ***Hit location*** (Figura 4), el estudiante podrá ir viendo cada caso en contexto donde aparece el verbo "*was*". Esto le permitirá identificar cuál de los ejemplos tiene error de conjugación.

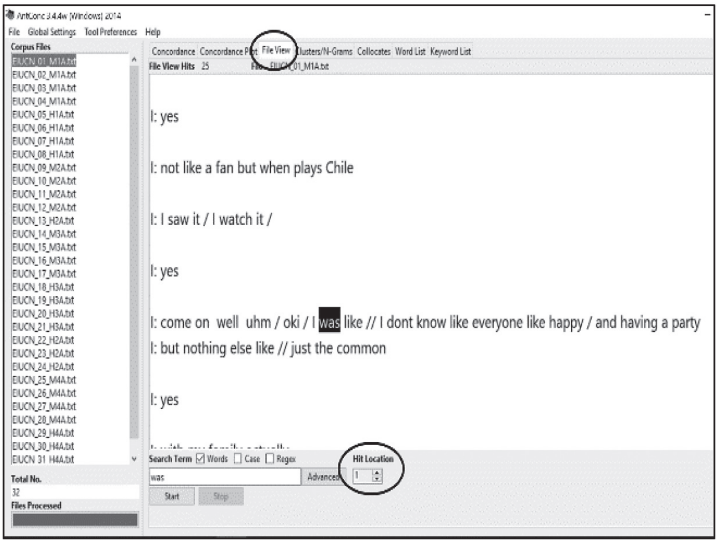

Figura 4. Ejemplo Hit location.

- *Etapa de análisis, discusión y sugerencias (Solo para estudiantes de tercero y cuarto año).*

Solicite a los estudiantes que determinen cuántos errores de conjugación existen por archivo. Este procedimiento les permitirá saber el número de veces que el verbo "was" aparece en cada archivo y en cuántos casos hay errores de conjugación. De este modo, se pueden establecer relaciones porcentuales. Un ejemplo sería el archivo 1. De los 25 casos del verbo "was" en el archivo, 10 tienen errores de conjugación; es decir, el 40%. En este ejercicio se pueden establecer comparaciones de los resultados por año y por sexo de los informantes, ya que el corpus posee esta información. Por tanto, los estudiantes pueden tener una idea general del manejo del verbo *to be* en pasado que poseen los informantes del corpus y sugerir actividades para reforzar el uso del verbo *was.*

- *Etapa de corrección (Solo para estudiantes de primero y segundo año).*

Una vez identificados los errores, extráigalos e imprímalos en una página con un espacio en blanco debajo de cada ejemplo para que los estudiantes corrijan la con-

jugación en cada caso tal y como se muestra en el ejemplo de la Figura 5.

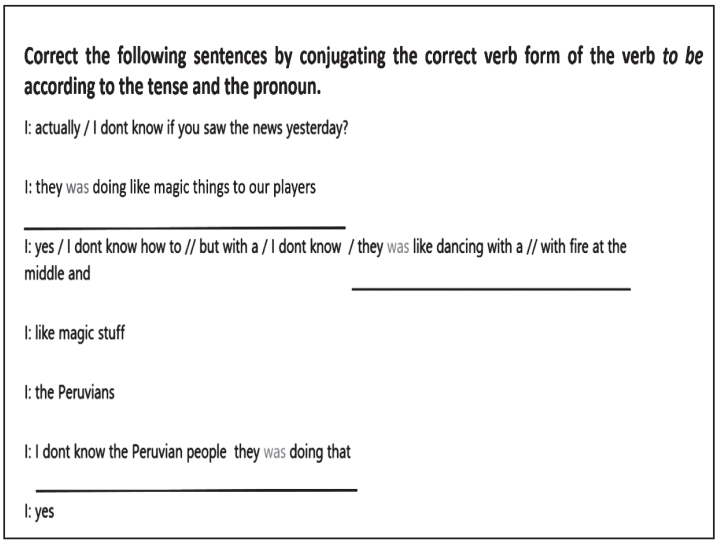

Figura 5. Ejemplo para la corrección por parte de los estudiantes.

5. CONSIDERACIONES FINALES

El ESOC-Chile representa un avance significativo en el ámbito del inglés como lengua extranjera por las siguientes tres razones. Primero, será de impacto en la enseñanza, principalmente en nuestra casa de estudios, ya que la misma puede estar basada en datos empíricos, los cuales orientarán: 1) la revisión y ajuste de contenidos en los programas de asignaturas, 2) la selección de estrategias didácticas al interior de la sala de clase y 3) la elaboración de materiales didácticos que se adapten de manera más precisa a las necesidades de nuestros estudiantes.

Segundo, el ESOC-Chile no solo brinda el escenario propicio para el estudio y comprensión de la producción oral, sino también para la incorporación de las tecnologías en los procesos de investigación y aprendizaje. Este avance se percibe en el entrenamiento de los estudiantes en el uso de programas de concordancias como el AntConc que, junto al uso de la base de datos, proporcionarán una oportunidad para el desarrollo y consolidación del aprendizaje autónomo, al mismo tiempo que despierta el interés por la investigación lingüística en la producción oral del idioma tanto a estudiantes como a académicos.

Finalmente, el presente corpus pasa a ser desde ahora en adelante un punto de referencia en Chile para la investigación en lingüística de corpus, el aprendizaje asistido por computadora y la creación de otros corpora a nivel nacional. El

mismo constituye la primera base sistematizada de datos oficiales y públicos en el país sobre la producción oral del inglés como lengua extranjera por parte de hispanohablantes.

REFERENCIAS

American National Corpus Project. (1990-). OANC: Open American National Corpus (15 millones de palabras 1990-2016). Disponible en http://www.anc.org/

Anthony, Laurence. (2019). AntConc (Versión 3.5.8) [Software Computarizado]. Tokyo, Japón: Waseda University. Disponible en http://www.laurenceanthony.net/software

Baker, Paul; Hardie, Andrew y McEnery, Tony. (2006). *A Glossary of Corpus Linguistics.* Findland: Edinburgh University Press.

Bennett, Gena. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers.* USA: The University of Michigan Press.

Bentivoglio, Paola y Malaver, Irania. (2006). La lingüística de corpus en Venezuela: Un nuevo proyecto. *Lingua Americana. 10*, pp. 37-46.

Bentivoglio, Paola y Malaver, Irania. (2012). Corpus Sociolingüístico de Caracas: PRESEEA Caracas 2004-2010 Hablantes de Instrucción Superior. *Boletín de Lingüística. XXIV* (37-38), pp. 144-180.

Cameron, Deborah. (1995). *Verbal Hygiene*. London: Routledge.

Coates, Jennifer. (1986). *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. London: Longman.

Council of Europe. (2014). *Common European Framework of References for Languages: Learning, Teaching, Assessment.* Cambridge University Press. Disponible en http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Davies, Mark. (2008-). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Disponible en https://corpus.byu.edu/coca/.

Eckert, Penelope. (1989). The Whole Woman: Sex and Gender Differences in Variation. *Language Variation and Change. 1*, pp. 245-67.

Eckert, Penelope y McConnell-Ginet, Sally. (2003). *Language and Gender*. Cambridge: Cambridge University Press.

Ehrlich, Susan, Meyerhoff, Miriam y Holmes, Janet. *(2014). The Handbook of Language, Gender, and Sexuality*. 2d ed. Malden, MA: Blackwell.

Holmes, Janet and Meyerhoff, Miriam. (1999). The Community of Practice: Theories and Methodologies in Language and Gender Research. *Language in Society*. *28*, pp. 173-183.

Hunston, Susan. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hymes, Dell. (1992). The concept of communicative competence revisited. En Martin Putz (Ed.) *Thirty Years of Linguistic Evolution: Studies in Honour of Rene Dirven on the Occasion of His Sixtieth Birthday*. Philadelphia / Amsterdam: John Benjamins, pp. 31-57

Instituto Cervantes. (2016). Corpus de aprendices de español (CAES). Disponible en http://www.cervantes.es/lengua_y_ensenanza/tecnologia_espanol/caes.htm

Jhons, Tim. (1991). Should You Be Persuaded – Two Samples of Data Driven Learning Materials. En Johns y King. (eds.). *Classroom Concordancing. ELR Journal, 4*. Birmingham: Birmingham University Press, pp. 1-16

Labov, William. (1994). *Principles of Linguistic Change, I: Internal Factors*. Oxford: Blackwell.

Labov, William. (2001). *Principles of Linguistic Change, II: Social Factors*. Oxford: Blackwell.

Lakoff, Robin. (1975). *Language and Woman's Place.* New York: Harper y Row.

Leech, Geoffrey. (1997). Teaching and Language Corpora: A Convergence. En Anne Wichmann; Steven Fligelstone; Tony McEnery y Gerry Knowles. (comps.). *Teaching and Language Corpora.* London: Longman, pp. 1-23.

McEnery, Tony y Wilson, Andrew. (2001). *Corpus Linguistics. An Introduction.* (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, Tony y Xiao, Richard. (2010). What Corpora Can Offer in Language Teaching and Learning. En Eli Hinkel, *Handbook of Research in Second Language Teaching and Learning.* Vol. 2. London y New York: Routledge, pp. 364-380.

Nesselhauf, Nadja. (2004). Learner Corpora and Their Potential for Language Teaching. En John Sinclair. *How to Use Corpora in Language Teaching*. Amsterdam: Benjamins, pp. 125-152.

Ortega, Maritza. (2014). Assessing Trainees' Oral Performance in a Chilean Teacher Training Program: A Corpus-Based Study. *Colombian Applied Linguistics Journal.* 16, pp. 10-16.

Oxford University Computing Services. (2007). The British National Corpus, versión 3 (Edición BNC XML). Disponible en http://www.natcorp.ox.ac.uk/

PRESEEA. (2014). *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América.* Alcalá de Henares: Universidad de Alcalá. Disponible en http://preseea.linguas.net

Reppen, Randi. (2010). *Using Corpora in the Language Classroom.* USA: Cambridge University Press.

Römer, Ute. (2009). Corpora and Language Teaching. En Anke Lüdeling y Merja Kytö. *Corpus Linguistics. An International Handbook. Vol. 1.* (pp. 112-129). Germany: Walter de Gruyter.

Silva, Carmen. (2001). *Sociolingüística y pragmática del español*. Washington DC:

Georgetown University Press.
Sinclair, John. (1980). COBUILD: Collins Birmingham University International Language Database. University of Birmingham, UK. Collins Publihers.
Stubbs, Michael. (2002). *Words and Phrases*. (2 ed). London, Blackwell.
Talbot, Mary. (1998). *Language and Gender: An Introduction*. Cambridge: Polity.
Tannen, Deborah. (1990). *You Just Don't Understand: Women and Men in Conversation*. New York: William Morrow.
TEI Consortium. (2016). TEI: Text Encoding Initiative. Disponible en http://www.tei-c.org/index.xml
Thi Ngoc, Doan. (3 de diciembre de 2013). Understanding the Relationship between Language and Gender. [Mensaje en un blog]. Disponible en http://gas.hoasen.edu.vn/en/gas-page/understanding-relationship-between-language-and-gender
Torruella, Joan y Llisterri, Joaquín. (1999). Diseño de corpus textuales y orales. En Blecua, Clavería, Sánchez y Torruella. *Filología e informática. Nuevas tecnologías en los estudios filológicos*. España: Editorial Milenio, pp. 45-77.
Trudgill, Peter. (1972). Sex, Covert Prestige and Linguistic Change in the Urban British English of Norwich. *Language in Society. 1*, pp. 179-195.
Zimmerman, Don y West, Candace. (1975). Sex Roles, Interruptions, and Silences in Conversation. En Barrie Thorne y Nancy Henley. *Language and Sex: Difference and Dominance*. Rowley, MA: Newbury House, pp. 105-29.

ANEXOS

Anexo 1. Tópicos para entrevistas

TOPIC 1 (5 min)

Option A: Chile and Its Soccer Team

1. Tell me your impressions of the result of Copa America.
2. What does the national soccer team mean to you?
3. Do you think Chile has a chance to win the next World Cup?
4. If you could think of a way to improve the national soccer team, what would it be?
5. Who is your national favorite player and why?

Option B: The Role of women in Chilean society

1. In your opinion, do Chilean women have the same opportunities to grow professionally as men in society?
2. Are you a(n) (anti)feminist? Why?
3. Do you agree having a female president has helped the role of women in Chile? Explain
4. In what ways do you think teenage pregnancy can be avoided?
5. Do you agree with abortion? Explain.

TOPIC 2 (5 min)

Education in Chile

1. What do you think of the quality of Chilean education at the moment?
2. Do you agree education should be free in Chile? Explain.
3. When compared to the education other countries in Latin America, the quality of education in Chile appears to be good and strong. Do you fear this quality may lower its standards if education goes free? Explain.
4. Would you agree that private education should be a choice for those who can pay for it?
5. In your opinion, what should be done to improve the quality of the education we have nowadays?

TOPIC 3 (5 min)

Pollution in Antofagasta

1. Do you think Antofagasta has a real problem of pollution? What is the evidence for this so called pollution? Explain.
2. How severe do you think pollution is in the city?
3. Who do you think should be held responsible for pollution in Antofagasta, the government, the mining companies or the citizens? Explain.
4. How can we eradicate pollution in our city?
5. In your opinion, how can education serve to fight pollution?

Anexo 2. Cabecera

<Trans audio_nombredelarchivo="EIUCN_01_M1A.mp3" xml:leng="ingles">
<Datos clave_texto="EIUCN_01_M11A" tipo_texto="entrevista_semidirigida">
<Corpus corpus="ESOC-Chile" ciudad="Antofagasta" pais="Chile"/>
<Grabacion resp_grab="Emily Noble" lugar="escuela de ingles" duracion="15'16""fecha_grab="10-13-2015" sistema="mp3"/>
<Transcripcion resp_trans="Rafael Arriquez" fecha_trans="11-01-2015" numero_palabras="1676"/>
<Revision num_rev="1" resp_rev="Chinger Zapata" fecha_rev="12-01-2015"/>
<Revision num_rev="2" resp_rev="Chinger Zapata" fecha_rev="06-15-2017"/>
</Datos><Hablantes>
<Hablante id="hab1" nombre="EIUCN_01_M1A" codigo_hab="I" sexo="mujer" nivel_edu="1A" estudios="pedagogia en ingles" profesion="estudiante" origen="Antofagasta" papel="informante"/>
<Hablante id="hab2" nombre="Emily Noble" codigo_hab="E" sexo="mujer" nivel_edu ="alto" estudios="filologia" profesion="docente" origen="Los Estados Unidos" papel="entrevistadora"/>
<Hablante id="hab3" nombre="Chinger Zapata" codigo_hab="A" sexo="hombre" nivel_edu="alto" estudios="filologia" profesion="docente" origen="Venezuela" papel="audiencia"/>
<Relaciones rel_ent_inf="profesor-alumno" rel_inf_aud="profesor-alumno"/> </Hablantes></Trans>[14]

[14] La plantilla de este texto de muestra sigue el modelo utilizado para el corpus PRESEEA. A la misma se le han realizado las modificaciones necesarias que se corresponden con la configuración del ESOC-Chile.

- Datos del propio archivo
nombre del archivo de audio (audio_filename): EIUCN_01_M1A.mp3
clave de texto (clave_texto): EIUCN_01_M1A.mp3 [código con formato general de ESOC-Chile: código del lugar – 5 caracteres –, número de entrevista, dado por el equipo, código de informante – sexo, grado de instrucción]
tipo de texto: entrevista semidirigida
corpus: ESOC-Chile
ciudad: Antofagasta
país: Chile
- Datos sobre la grabación de la entrevista
responsable de grabación (resp_grab): Emily Noble
lugar de grabación: Escuela de Inglés
duración: 15'16"
fecha de grabación (fecha_grab): 10-13-2015 [las fechas que aparecen en la cabecera tienen el formato mm-dd-aa]
sistema: mp3 [tipo de archivo de grabación original]
- Datos sobre la transcripción y revisión de la entrevista
responsable de transcripción (resp_trans): Rafael Arríquez
fecha de transcripción (fecha_trans): 11-01-2015 [las fechas que aparecen en la cabecera tiene el formato mm-dd-aa]
extensión (numero_palabras): 1676 [el número de palabras corresponde al texto transcrito, excluidas la cabecera completa y las etiquetas]
revisión 1 (resp_rev): Chinger Zapata
fecha revisión 1 (fecha_rev): 12-01-2015 [las fechas que aparecen en la cabecera tienen el formato mm-dd-aa]
revisión 2 (resp_rev): Chinger Zapata
fecha revisión 2 (fecha_rev): 06-15-2017 [las fechas que aparecen en la cabecera han de tener el formato mm-dd-aa]
-Datos sobre los hablantes participantes en la entrevista
nombre de informante (nombre): EIUCN_01_M1A [se utilizó el mismo código que figura en el campo "clave de texto"; en este caso, cada equipo podría añadir, tras un nuevo guion bajo, otro código que permitiera su identificación con fines particulares]
código hablante (código_hab): I [los códigos de hablante son I (Informante), E (Entrevistador) y A (Audiencia)]
sexo: mujer [hombre | mujer]
nivel educativo (nivel_edu): 1A [1 Año|2 Año |3 Año |4 Año]
estudios: pedagogía en inglés
profesión: estudiante
origen: Antofagasta
papel: informante
nombre de entrevistador (nombre): Emily Noble

código hablante (codigo_hab): E
sexo: mujer [hombre | mujer]
nivel educativo (nivel_edu): alto [bajo | medio | alto]
estudios: filología
profesión: profesora
origen: Los Estados Unidos
papel: entrevistador
nombre de audiencia 1: Chinger Zapata [puede haber más de una persona (o interlocutor) como audiencia; los datos que se desconocen reciben el valor "desc" (desconocido); si no existe audiencia, se elimina de la cabecera el segmento correspondiente al id="hab3"]
código hablante (código_hab): A1 [en caso de haber más de un hablante como "audiencia", pueden utilizarse los códigos A2, A3, …]
sexo: hombre [hombre | mujer]
nivel educativo (nivel_edu): alto [bajo | medio | alto | desc]
estudios: filología [también podría ser "desc"]
profesión: profesor [también podría ser "desc"]
origen: Venezuela [también podría ser "desc"]
papel: audiencia
relación E-I (rel_entre_inf): profesor-estudiante [relación entre entrevistador e informante: conocidos | desconocidos]
relación I-A1 (rel_inf_aud1): profesor-estudiante [relación entre informante y audiencia: conocidos | desconocidos | no] [si no existe A1, debe anotarse la opción "no"]
relación E-A1 (rel_entre_aud1): conocidos [relación entre entrevistador y audiencia: conocidos | desconocidos | no] [si no existe A1, debe anotarse la opción "no"]

Anexo 3. Marcas y etiquetas presentes en el ESOC-Chile

ORTOGRAFÍA Y PUNTUACIÓN

! Enunciados exclamativos
? Enunciados interrogativos
/ Pausa mínima
// Pausa
: Tras código de hablante (I: E: A1:)
Mayúsculas: Inicial de nombres propios, siglas y el pronombre de primera persona del singular para el inglés (I).
Elementos cuasi-léxicos funcionales: Interjecciones; apoyos. Escritura ortográfica (alas, ah, aw, dear, eh, eek, er, hey, hm, hmm, huh, mmm, nah, oh, oops, phew, uh, uhm, uh-huh, uh-hum, well, entre otras.)
Onomatopeyas: Escritura ortográfica (zas, bum, plas)

ETIQUETADO DE RUIDOS

<ruido = " "**/>** Ruido, con especificación de tipo (ejemplo: <ruido ="chasquido boca"/>) I

<ruido_fondo> </ruido_fondo> Ruido continuo de fondo AD

<risas = " "**/>** Risas, con especificación de emisor/es (ejemplo: <risas ="E"/>, <risas = "todos"/>) I

<entre_risas> </entre_risas> Risas simultáneas con el habla AD

<registro_defectuoso> </registro_defectuoso> Fragmento de la grabación de mala calidad AD

<interrupción_de_grabación/> Interrupción de la grabación I

ETIQUETADO FÓNICO

<énfasis> </énfasis> Fragmento con pronunciación claramente enfática AD

<alargamiento/> Alargamiento de sonido D (sin espacios)

<silencio/> Silencio de cinco segundos o más I

<palabra_cortada/> Palabra cortada D

<vacilación/> Vacilación; titubeo breve I

<sic> </sic> No es descuido de transcripción AD

<ininteligible/> Fragmento ininteligible I

ETIQUETADO LÉXICO

<término> </término> Lexía claramente usada como uso especializado AD

<extranjero> </extranjero> Extranjerismo (excepto usos de la L2 del hablante) AD

<siglas = [ ]**> </siglas>** Siglas; incluye pronunciación AD

ETIQUETADO DE DINÁMICA DISCURSIVA

<cita> </cita> Cita, estilo directo AD

<simultáneo> </simultáneo> Solapamiento (traslape). También se usa en turnos de apoyo, si fuera necesario AD

ETIQUETADO DE LENGUA

<lengua = " "**> </lengua>** Cambio de lengua (léxico, oracional, …), especialmente L1 del hablante, con indicación desarrollada de lengua (ejemplo <lengua = "español"> </lengua> AD

ETIQUETADO DE TRANSCRIPCIÓN

<transcripción_dudosa> </transcripción_dudosa> Transcripción dudosa para transcriptor y revisores AD

<tiempo = " "**/>** Anotación de minuto y segundo de grabación. (ejemplo: <tiempo = "02:45"/>) I

<observación_complementaria = " "**/>** Observación complementaria I

El siguiente ejemplo muestra un breve fragmento de texto de una de las entrevistas con etiquetas.

E: uhm <vacilación/>m/ Im just gonna ask you some questions and you respond with your own

I: oki

E: opinions / uhm <vacilación/>m <vacilación/>/ ok / so / do you watch / do you watch soccer at all?

I: yes

E: youre a soccer fan?

I: not like a fan but when plays Chile

Anexo 4. Texto sin etiquetas

<Datos
Sexo: mujer
Profesión: estudiante
Nivel educacional: primer año
Entrevista: 01
Fecha de grabación: 13-10-2015
Tipos de palabras: 395
Número de palabras: 1676>

E: uhm / Im just gonna ask you some questions and you respond with your own
I: oki
E: opinions / uhm / ok / so / do you watch / do you watch soccer at all?
I: yes
E: youre a a soccer fan?
I: not like a fan but when plays Chile