Lingüística y Literatura
ISSN: 0120-5587
ISSN: 2422-3174
Universidad de Antioquia

Pardo Rodríguez, María Victoria

LA CONSTRUCCIÓN DE UN CORPUS COMPUTARIZADO DE ERRORES ESCRITOS CON TEXTOS DE ESTUDIANTES UNIVERSITARIOS EN COLOMBIA

Lingüística y Literatura, núm. 78, 2020, Julio-Diciembre, pp. 35-54
Universidad de Antioquia

DOI: https://doi.org/10.7440/res64.2018.03

Disponible en: https://www.redalyc.org/articulo.oa?id=476569499002

LA CONSTRUCCIÓN DE UN CORPUS COMPUTARIZADO DE ERRORES ESCRITOS CON TEXTOS DE ESTUDIANTES UNIVERSITARIOS EN COLOMBIA

María Victoria Pardo Rodríguez
Universidad del Norte (Colombia) / Universidad de Antioquia (Colombia)
mvpardo@uninorte.edu.co; victoria.pardo@udea.edu.co; mpardoenudea@gmail.com

Recibido: 24/11/2019 - **Aprobado:** 04/02/2020
DOI: doi.org/10.17533/udea.lyl.n78a02

Resumen: El presente artículo describe las etapas de una investigación doctoral en la construcción de un corpus computacional a partir de textos escritos por estudiantes de inglés como lengua extranjera. La compilación de los datos se realizó conforme a los procedimientos de la lingüística de corpus computacional (McEnery & Hardie, 2011). El corpus permitió determinar los errores más frecuentes de los estudiantes con el fin de conocer el nivel de interlengua y mejorar las prácticas docentes para el aprendizaje.
Palabras clave: corpus lingüístico computacional; corpus de aprendientes; análisis de errores; interlengua; anotación de errores.

THE CONSTRUCTION OF A COMPUTERIZED CORPUS OF WRITTEN ERRORS WITH TEXTS OF UNIVERSITY STUDENTS IN COLOMBIA

Abstract: This article describes the stages of doctoral research in the construction of a computational corpus from texts written by students of English as a foreign language. The compilation of the data was carried out according to the procedures of the computational corpus linguistics (McEnery & Hardie, 2011). The corpus allowed determining the most frequent errors of students in order to know the level of interlingua and to improve teaching practices for learning.
Key words: computational corpus linguistics; learner corpus; error analysis; interlanguage; error annotation.

1. Introducción

El término «corpus» usado en un contexto lingüístico moderno tiene varias connotaciones en relación con su tamaño, representatividad y formato electrónico. Existe un consenso general sobre las características que debe tener un corpus digital: debe estar formado por textos auténticos en formato electrónico —legible y a máquina— y además, debe incluir ejemplos que representan una lengua o una variedad en particular (McEnery, Xiao, & Tono, 2006). Se define al corpus como una colección de textos o de partes de textos que han sido puestos en formato electrónico y seleccionados de acuerdo a determinados criterios para representar una lengua o una variedad de lengua y que es usado como fuente de datos para investigación lingüística. (Sinclair, 2005), (Crystal, 2008), (McEnery & Hardie, 2011). La lingüística de corpus (LC) es una metodología que permite realizar investigaciones confiables utilizando grandes cantidades de datos. Gracias al avance en el campo de las tecnologías de la información (IT), la LC se rige por pasos y procedimientos definidos para realizar la recopilación de datos y el análisis científico de éstos. Sus procedimientos exploran y describen el lenguaje objetivamente dejando atrás la especulación subjetiva.

Existen diferentes tipos de corpus de acuerdo a sus funciones. Así por ejemplo, en un corpus general se pueden encontrar varios tipos de textos producidos en uno o en diferentes países. También se pueden encontrar los corpus de muestra o monitor, los cuales son diacrónicos y contienen ejemplos de variedad de textos que permiten realizar análisis sobre los cambios y evolución de una lengua. Dentro de los corpus especializados que compilan textos con el fin de realizar descripciones e investigaciones específicas de una lengua, encontramos los corpus de aprendientes. Estos corpus se definen como «colecciones electrónicas de datos naturales o casi naturales producidos por estudiantes extranjeros o de segunda lengua (L2) y reunidos según criterios de diseño explícitos» (Granger, 2002, p. 7), (Gilquin, 2015, p. 1). De acuerdo a Granger, (2002) Los corpus electrónicos de aprendientes deben ser codificados de manera estandarizada, homogénea y documentados de acuerdo a su origen y procedencia. Los datos de los corpus de aprendientes se diferencian de otros corpus en dos aspectos: primero, están totalmente en formato digital, lo

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

cual permite su estudio usando diferentes herramientas lingüísticas de software para obtener un análisis rápido y eficiente y; segundo, la cantidad de datos recopilados es considerable, por lo cual se constituye en una base confiable para describir y ejemplificar la lengua de los aprendientes (Granger, 2003). Estos corpus proveen datos que dan a conocer los mecanismos de adquisición de lenguas, ya sean éstas foráneas o segundas lenguas (p. 5).

Los corpus de aprendientes surgieron al final de la década de 1980 como una forma científica válida para analizar la producción oral o escrita de los estudiantes. Estos corpus tienen las mismas características atribuidas a otros corpus; sin embargo, los datos en este caso se obtienen de la producción oral o escrita de los estudiantes colocada en un formato electrónico o multimodal (audiovisuales) para ser analizada por medio de software especializado. Los corpus de aprendientes son colecciones auténticas de uso del lenguaje contextualizado que contiene muestras de la adquisición de un segundo idioma o de un idioma extranjero.

La lingüística de corpus computacional (LCC) se encarga del diseño de herramientas informáticas para lograr una mejor comprensión del lenguaje y un análisis confiable, eficiente y automático (Bolaños, 2015, p. 31). En Colombia la LCC es relativamente nueva, ya que los estudios lingüísticos se han basado en corpus sin el uso de herramientas de la lingüística computacional. Esta situación se puede evidenciar en los trabajos más recientes basados en corpus, (Londoño, 2008), (Parada, Ruiz, & Sánchez, 2017). Lo anterior es evidencia de la necesidad de ilustrar a los lingüistas investigadores colombianos y de la región sobre la importancia del diseño de corpus computacionales de aprendientes. El presente artículo se configura como un primer paso para llenar este vacío.

2. Marco teórico

En esta sección se describen los criterios y principios en la elaboración de un corpus computacional de aprendientes y se hace un recuento de los últimos estudios con este tipo de corpus en Latinoamérica.

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

Los corpus electrónicos de aprendientes (CEA) deben cumplir determinados principios de diseño ya que deben ser compilados teniendo en cuenta criterios específicos de la lingüística de corpus, pero además pertenecen a un tipo de corpus específico con datos obtenidos de la producción oral o escrita de los estudiantes de lenguas. La figura 1 proporciona algunas pautas que se deben tener en cuenta en el diseño de un corpus de aprendientes de acuerdo a Granger (2002).

Aprendiz	**Configuración de tareas**
• Contexto de aprendizaje • Lengua materna • Otras lenguas foráneas • Nivel de competencia • […]	• Límite de tiempo • Uso de herramientas de referencia • Tipo de prueba • Audiencia/interlocutor • […]

Figura 1. *Criterios de diseño específicos de los CEA*. Fuente: Granger (2002).

Al diseñar este tipo de corpus se deben tener en cuenta, además de los aspectos señalados, las variables de los aprendientes (edad, ambiente de aprendizaje, nivel de competencia, género, lengua materna, región, exposición a la segunda lengua) y las variables propias de la tarea o actividad encomendada (medio, campo, género, extensión, tema, herramientas que se pueden usar) (Granger, 2003). Todos estos aspectos conforman la información relevante al compilar un corpus de estudiantes de lengua.

Una de las razones para compilar un corpus es analizar el nivel de interlengua entre los estudiantes mediante el análisis de errores (AE). El AE permite, entre otras: clasificar y analizar los errores, determinar los factores que contribuyen al éxito o fracaso del aprendizaje y metodologías de enseñanza, conocer las causas y origen de los errores, emitir juicios sobre gramática y validez de normas lingüísticas en determinada lengua (Vázquez, 2009). Con el auge de las nuevas tecnologías surge el análisis de errores asistido por computador (del inglés *Computer-Aided-Error Analysis* - CEA) el cual basa la investigación en corpus electrónicos de aprendientes teniendo en cuenta los procedimientos de la lingüística de corpus (LC) para identificar, clasificar y describir los errores. Los resultados obtenidos por medio del CEA contribuyen con innovación en el procesamiento de los datos que son analizados por medio de software para mejorar las competencias y

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

habilidades lingüísticas y para diseñar materiales más acordes y tareas más significativas en el proceso de aprendizaje de lenguas extranjeras. Los diferentes análisis han permitido desarrollar modelos de tratamiento de errores que buscan profundizar en la identificación y reflexión de los mismos, con el fin de sensibilizar y orientar a los aprendientes hacia la autocorrección y fomentar su investigación (Alexopoulou, 2005). Por otra parte, diferentes autores, (Vázquez, 2009; Alba Quiñones, 2009) proponen taxonomías de errores y pruebas con el fin de identificarlos.

Se puede afirmar entonces que los corpus de aprendientes se recopilan cumpliendo determinados criterios y que uno de los fines de su existencia es realizar análisis de errores. Una vez que el corpus ha sido compilado siguiendo los principios de la lingüística de corpus y conforme a los corpus de aprendientes, el siguiente paso es iniciar la anotación del corpus.

2.1. Sistema de anotación

Los corpus de aprendientes, como cualquier otro tipo de corpus, comienzan como textos en bruto de versiones escritas en formato electrónico o textos transcritos de la producción oral de los estudiantes. Un corpus de aprendientes es más productivo cuando ha sido anotado porque adquiere un valor agregado. La anotación es la práctica de agregar información interpretativa y lingüística a un corpus electrónico de datos del lenguaje hablado o escrito (Leech, 2005, p. 25). La información agregada viene en forma de etiquetas, que son entidades individuales agregadas a una parte o partes del discurso. Las etiquetas son únicas y pueden identificar características del corpus analizado. Existen diversos tipos de esquemas de anotación y requieren diferentes etiquetas según el objetivo del investigador; por ejemplo, la lingüística descriptiva utiliza etiquetas para identificar partes del habla para así obtener una anotación gramatical en un corpus. Otro ejemplo es la anotación semántica, que requiere asignar un campo semántico a una palabra y se usa para refinar búsquedas y clasificaciones de acuerdo con el propósito de la investigación. Un corpus de aprendientes también puede tener anotaciones de errores para clasificarlos y analizarlos. El corpus recopilado durante la presente investigación fue etiquetado de

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

acuerdo a ocho categorías de errores que en total suman 56 tipos de errores de acuerdo al Manual de Etiquetado de Errores de la Universidad de Lovaina (Dagneaux *et al.*, 2005). Todas las categorías y etiquetas de error pueden observarse en el anexo 1 del presente artículo.

2.2. Alineación del corpus

En el caso particular de un corpus de errores, la alineación se refiere a la identificación de patrones de error identificados mediante las etiquetas de error. Para realizar este procedimiento se utiliza un software especializado que agrupa en listas los tipos de errores.

2.3. Estadística de los errores en un corpus de aprendientes

Este paso se refiere a la obtención de estadísticas de los patrones de error o de los patrones lingüísticos buscados objeto de la investigación. En el caso del presente corpus, se obtuvieron las estadísticas de los errores por categoría y por tipo.

2.4. Estado de la lingüística de corpus computacional en Colombia y en Latinoamérica

A pesar de que la lingüística de corpus computacional ha tenido su auge desde los años 80 en Colombia, apenas se empieza a considerar para la realización de análisis lingüísticos. La lingüística de corpus computacional se ha venido posicionando en países como México donde se han realizado trabajos con corpus computacionales de aprendientes (López, 2008). En su mayoría los trabajos reportados por AELINCO *Research in Corpus Linguistics* se refieren a investigaciones con corpus de aprendientes recolectados en Europa, (Hernandez, 2013; Szabó, 2013; Crespo, 2016) entre otros. Desde hace varios años Chile se configura como el país que más ha aprovechado este método de compilación siendo el más prolífico en este tipo de investigaciones a nivel latinoamericano. Esto se puede ver en el auge de trabajos relacionados con la lengua materna y en la enseñanza del inglés como lengua extranjera (ILE): (Garrido, 2012; Cabrera, Elejalde, & Vine, 2014; Ortega, 2014; Saavedra

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

& Campos, 2018) entre otros. Colombia es un país con mucho potencial para el uso de corpus computacionales de aprendientes por la necesidad de analizar grandes cantidades de datos provenientes de la interlengua de los estudiantes. Al revisar algunos de las últimos trabajos con corpus de aprendientes en Colombia no se encuentran corpus computacionales (Vásquez, 2008; Parada *et al.*, 2017) lo cual dificulta llevar a cabo análisis comparativos y contribuye al estancamiento en esta área, de ahí la importancia del presente análisis.

Es necesario, pues, que la lingüística de corpus computacional se posicione en Colombia para que a futuro se puedan hacer análisis a gran escala usando herramientas de software que garanticen la confiabilidad en el análisis de grandes cantidades de datos. El presente trabajo busca abrir las puertas a este fascinante mundo de la lingüística computacional.

3. Metodología: Análisis de errores y lingüística de corpus

El problema de investigación que se propone abordar el presente estudio es el vacío que existe en Colombia y en gran parte de Latinoamérica sobre literatura referente a la construcción de corpus computacionales.

La presente investigación se sustenta en la metodología de la lingüística de corpus computacional en la compilación del corpus de aprendientes en formato electrónico, (Granger, 2002; Granger, 2003; Gilquin, 2015).

A continuación, se describirán los pasos seguidos para construir un corpus computarizado de estudiantes de inglés como idioma extranjero a nivel universitario.

3.1. Recolección del corpus

El corpus de textos lo conforman 515 textos escritos: 112 son ensayos de comparación y contraste y 403 son párrafos de opinión. Las siguientes son las etapas seguidas en la recolección y alistamiento del corpus de aprendientes:

—Aplicación de encuesta y registro de participación: Este instrumento estaba conformado por 27 preguntas divididas en tres secciones: perfil del estudiante, perfil

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

académico y aspectos socioculturales. Se buscaba determinar datos tales como la edad, el estrato socio-económico, la lengua materna, el manejo de otras lenguas etc.

—**Compilación del corpus:** El corpus de la presente investigación se compiló en la Universidad del Norte, en Barranquilla, (Colombia) durante el segundo semestre del año 2015. En total, 2 088 estudiantes universitarios de pregrado en diferentes carreras estaban inscritos en los cursos de inglés como lengua extranjera en los niveles B1.1 a B2.3, de acuerdo a la clasificación del Marco Común Europeo de Referencia - MCER. De esa población, 515 estudiantes firmaron un formulario de consentimiento para participar en esta investigación.

Por medio de instrumentos con preguntas sobre diferentes temas de la vida cotidiana se recopilaron 403 párrafos de opinión y 112 ensayos de comparación y contraste. Los estudiantes iniciaron el proceso de escritura desarrollado durante varias clases, en las que recibían comentarios y sugerencias de cada docente. Los trabajos recolectados para el presente análisis corresponden al ejercicio final de escritura con tiempo limitado. El análisis se centró en los niveles B1.1 a B2.3 de acuerdo al MCER (Council of Europe, 2001). La duración del curso fue de un semestre (16 semanas) para un total de 64 horas de exposición a la lengua con una intensidad de cuatro horas de clase semanales. Todos los cursos se desarrollaron como seminarios con un producto por clase que era evaluado, corregido y recibía comentarios. Desde los niveles B1.1 a B1.3, los estudiantes entregaron tanto sus borradores como el trabajo final escritos a mano, pues los laboratorios de informática se reservan para los estudiantes avanzados que tienen mayor producción escrita. Desde el nivel B2.1 hasta el B2.3, los borradores y el trabajo final se entregaron en archivo electrónico. En total, el corpus compilado de estos estudiantes tenía 149 325 tokens, 12 164 tipos y 12 337 lemas. En la Tabla 1 se presenta la distribución de los alumnos según los niveles del MCER.

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

Levels									
U. Norte Levels	**Introductory**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
MCER Levels	**A1.1**	**A2.1**	**A2.2**	**B1.1**	**B1.2**	**B1.3**	**B2.1**	**B2.2**	**B2.3**
# Students	110	496	439	409	325	356	377	355	286
				Pre-Intermediate	Intermediate	Intermediate II		Upper Intermediate	

Tabla 1. *Distribución de estudiantes registrados en niveles B1.1 a B2.3*

En cada nivel, desde B1.1 hasta B2.3, los estudiantes recibieron una lista de posibles temas para escribir sus composiciones. Este grupo de estudiantes no tuvo acceso a dispositivos electrónicos. Los estudiantes del nivel B2.1 hasta el B2.3 realizaron un ensayo de comparación y contraste. En este caso, los estudiantes tuvieron acceso a diccionarios en línea y a fuentes secundarias a través de las bases de datos de la plataforma de la universidad o a través de documentos sugeridos o proporcionados por sus maestros durante el semestre. Se considera que al ser los ensayos un tipo de texto crítico es relativamente difícil cortar y pegar información a sus textos sin quedar en evidencia en los programas detectores de plagio usados en la institución. En todos los casos, los estudiantes tuvieron un tiempo limitado de una clase de dos horas para realizar la tarea final recopilada para la presente investigación.

—Técnicas de estimulación usadas para generar producción escrita de los estudiantes: Estas técnicas buscan generar un tipo de respuesta abierta en la que se puede establecer el estado de la interlengua de los estudiantes (Corder, 1981, p. 61). Los textos son el resultado de una serie de actividades que guían al estudiante a generar una respuesta escrita. Tanto en los párrafos de opinión como en los ensayos los estudiantes recibían posibles temas para desarrollar. En la Figura 2 se esbozan los pasos realizados en el proceso de escritura de los párrafos, el cual fue similar aunque con un resultado esperado diferente en el caso de los ensayos.

Paso 1:
Se enuncia la actividad y lo que se espera que hagan los estudiantes

What am I going to do?
You are going to ***write an opinion paragraph***. Your paragraph should include the

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

following (see rubric):

a) include at least one compound sentence (a sentence that joins two independent clauses using a coordinating conjunction like *and, or, but, so*)

Paso 2:

El estudiante tiene la opción de escoger entre tres diferentes temas.

Option One	**Option Two**	**Option Three**
Are commercials on TV honest? Provide several good examples to support your answer.	Should con men and women who are caught in the act of fraud be punished severely? Explain.	In your opinion, do tourists pose a danger to the environment? Support your argument with several good examples.

Paso 3:

En este caso los estudiantes tenían un corto periodo de tiempo para una lluvia de ideas.

Step Three: Brainstorming (5 Minutes)

a) Take a few minutes to think about which option you want to write about.

b) Then, think about what you are going to write. Make a list of your ideas in the space provided.

Paso 4:

Se pide a los estudiantes hacer un esquema con las partes relevantes del escrito

Step Four: Write an OUTLINE (10 minutes)

a) Begin by thinking about a strong **opinion** statement. This is your TOPIC sentence. Write it down in the space provided for the topic sentence. It should be brief and to the point.

b) Now write at least 2 or 3 good reasons that answer why your opinion makes sense. Write your reasons in the space provided. Your reasons should be brief and to the point. Include examples or explanation for each reason.

c) Finally, write down a conclusion that restates your topic sentence in a different way. Make sure to use different vocabulary to express your conclusion.

Paso 5:

Se inicia en este paso la escritura en forma de párrafo.

Step Five: Write a First Draft (15 minutes)

a) Rewrite your OUTLINE as a paragraph on the space provided. Remember to **double-space** so your work is clear.

b) When you finish, read your paragraph and go through the **checklist** (5 minutes). Once your checklist is complete, it is time to write your final draft.

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

Paso 6:
El párrafo final se debe entregar junto con todos los pasos anteriores.

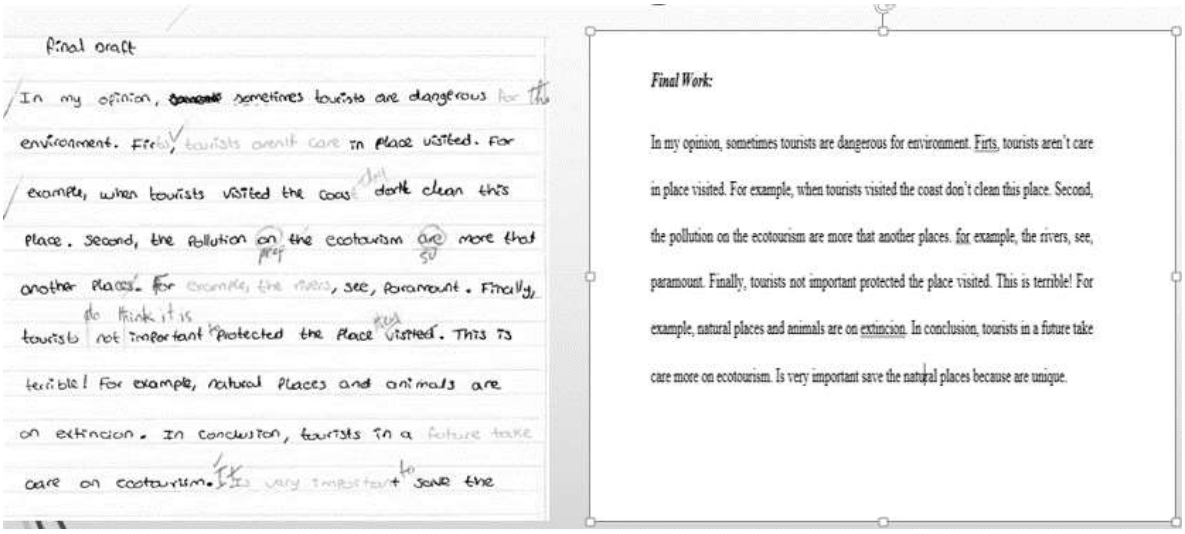
Step six: Rewrite Final Draft (10 minutes)

a) Rewrite your paragraph and hand ALL your work in before you leave. **Do not forget to include your name.**

Figura 2. *Pasos realizados en el proceso de escritura*

Luego de la compilación de los textos, los archivos fueron sometidos a procesos diferentes porque estaban en formatos diferentes. Por ejemplo, en los niveles B1.1 a B2.1 el trabajo final de los estudiantes fue escrito a mano, el proceso comenzó con el escaneo de los textos seguido de su transcripción. Después de transcribir todos los textos, asegurándose de que no se hicieron adiciones o eliminación de palabras, los archivos se verificaron dos veces para asegurar que eran exactamente como el original. A continuación, se convirtieron a formato txt para llevar a cabo la anotación de los errores. Los estudiantes de los niveles B2.2 a B2.3 hicieron la versión digital directamente, por lo tanto, esos textos se convirtieron inmediatamente en el mismo formato que los anteriores para el etiquetador de errores. Un total de 403 textos escritos a mano fueron compilados y transcritos en archivos de Microsoft Word. Después de la preparación anterior, todos los archivos estaban listos para comenzar con el proceso de etiquetado de errores. La Figura 3 muestra un ejemplo de un archivo escrito a mano que se transcribió a un archivo de Microsoft Word.

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54



final draft

In my opinion, ~~second~~ sometimes tourists are dangerous for the environment. Firts, tourists aren't care in place visited. For example, when tourists visited the coast don't clean this place. Second, the pollution on the ecotourism are more that another places. for example, the rivers, see, paramount. Finally, tourists not important protected the place visited. This is terrible! For example, natural places and animals are on extincion. In conclusion, tourists in a future take care on ecotourism. It is very important to save the

Final Work:

In my opinion, sometimes tourists are dangerous for environment. Firts, tourists aren't care in place visited. For example, when tourists visited the coast don't clean this place. Second, the pollution on the ecotourism are more that another places. for example, the rivers, see, paramount. Finally, tourists not important protected the place visited. This is terrible! For example, natural places and animals are on extincion. In conclusion, tourists in a future take care more on ecotourism. Is very important save the natural places because are unique.

Figura 3. *Archivo manuscrito y luego transcrito a formato digital en Microsoft Word*

—**Etiquetado del corpus**: El etiquetado del corpus hace parte de la descripción de los errores. Es la primera etapa en la que se reconoce su existencia. La identificación y clasificación de los errores se hizo de acuerdo al Manual de Etiquetado de Errores versión 1.2 (*Error Tagging Manual versión 1.2*) de la Universidad de Lovaina (Dagneaux *et al.*, 2005). El etiquetador, además de describir el nivel lingüístico y el tipo de errores, presenta una explicación de los posibles procesos que subyacen a las etapas de la interlengua. El manual de etiquetado distingue entre ocho categorías y casi cada categoría se divide en subcategorías para un total de 56 etiquetas de error. En cada caso, la primera letra de la etiqueta muestra la categoría de error. La Figura 4 presenta las categorías de error utilizadas en el presente estudio. Cada categoría de error representó varios tipos de error.

Categoría	**Código. (letra que representa la categoría de error)**
1. Errores de Forma	F
2. Errores gramaticales, v.g. errores que rompen las reglas generales de la gramática inglesa.	G
3. Errores léxico-gramaticales, v.g. errores donde las propiedades morfosintácticas de una palabra han	X

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

sido violadas.	
4. Errores de léxico, v.g. errores que involucran las propiedades semánticas de algunas palabras y frases.	L
5. Palabra redundante o faltante y otros errores que involucran el orden de las palabras.	W
6. Errores de puntuación.	Q
7. Errores de estilo.	S
8. Infelicities, no son propiamente errores sino expresiones poco o no aceptadas pragmáticamente en un contexto.	Z

Figura 4. *Categorías de errores con sus etiquetas*. Fuente: Dagneaux *et al.* (2005).

Cada anotación se hizo siguiendo las ocho categorías de la versión 1.2 del etiquetador referenciado anteriormente (Dagneaux *et al.*, 2005). Este etiquetador es el sistema más apropiado porque se especializa en la anotación de errores de corpus de aprendientes de lengua. Las etiquetas incluyen la descripción de los errores más comunes de los estudiantes de inglés como lengua extranjera y cuya lengua materna es el español. Este etiquetador es ampliamente conocido por los analistas de errores, por lo que su uso garantiza que en el futuro se pueden hacer trabajos comparables para tener la validación externa de los datos. La Figura 5 muestra un ejemplo de cómo se hizo la etiquetación de errores. Cuando se detectó un error, la etiqueta se colocó justo antes del error y la corrección sigue el error entre dos signos de dólar: $ corrección $ como lo indica el manual:

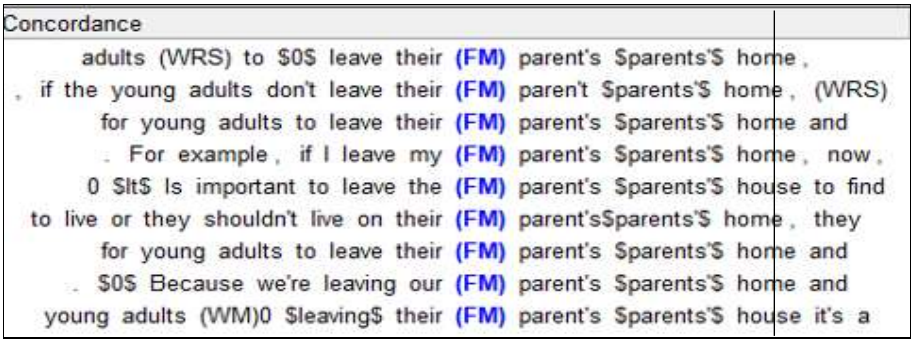
Ejemplo:
Nowadays we have seen (GADJN) differents $different$ (este error corresponde a la categoría Gramática y se refiere a la pluralización de un adjetivo en inglés).

Figura 5. *Ejemplo de etiquetado de un error*

Al terminar la etapa de etiquetado de errores en todo el corpus y la doble revisión de su consistencia se procedió con la alineación del corpus.

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

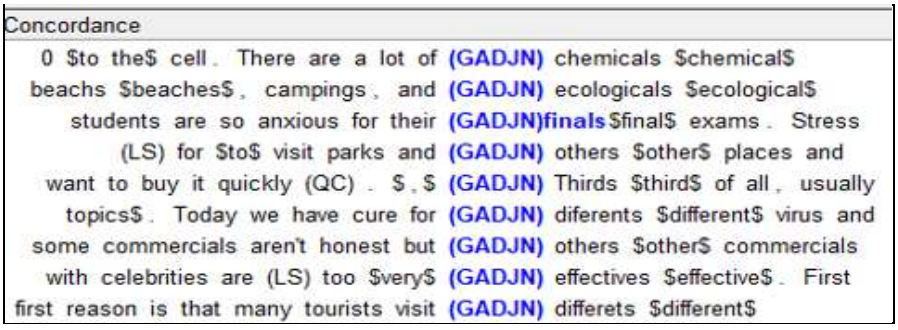
—Extracción y alineación del corpus por tipo de error: El siguiente paso después de hacer el etiquetado de errores es la extracción y alineación del corpus. Este proceso se realiza por medio de un software de extracción que busca las etiquetas y las agrupa de acuerdo a cada tipo de error. Las etiquetas se extraen dentro de un contexto que permite un análisis adecuado. La alineación del corpus se hizo por medio del software Word Smith (Scott, 2008) y LancsBox (Brezina, McEnery, & Wattam, 2015) que permiten la identificación de patrones de lengua y la obtención de estadísticas de los datos con sus respectivas gráficas. A continuación, se puede apreciar en las figuras 6, 7 y 8 la incurrencia en tres diferentes tipos de errores en sus respectivos contextos.


Concordance

adults (WRS) to 0 leave their (FM) parent's $parents'$ home ,
, if the young adults don't leave their (FM) paren't $parents'$ home , (WRS)
for young adults to leave their (FM) parent's $parents'$ home and
. For example , if I leave my (FM) parent's $parents'$ home , now ,
0 It Is important to leave the (FM) parent's $parents'$ house to find
to live or they shouldn't live on their (FM) parent's$parents'$ home , they
for young adults to leave their (FM) parent's $parents'$ home and
. 0 Because we're leaving our (FM) parent's $parents'$ home and
young adults (WM)0 $leaving$ their (FM) parent's $parents'$ house it's a

Figura 6. *Alineación de error FM (Error de Forma, Morfología)*

La Figura 6 muestra al error FM (Forma, Morfología) en nueve ocurrencias diferentes y con sus respectivos contextos. En este caso se detectan los errores en el morfema genitivo sajón.


Concordance

0 $to the$ cell . There are a lot of (GADJN) chemicals $chemical$
beachs $beaches$, campings , and (GADJN) ecologicals $ecological$
students are so anxious for their (GADJN)finals $final$ exams . Stress
(LS) for to visit parks and (GADJN) others $other$ places and
want to buy it quickly (QC) . $, $ (GADJN) Thirds $third$ of all , usually
topics$. Today we have cure for (GADJN) diferents $different$ virus and
some commercials aren't honest but (GADJN) others $other$ commercials
with celebrities are (LS) too $very$ (GADJN) effectives $effective$. First
first reason is that many tourists visit (GADJN) differets $different$

Figura 7. *Alineación de error GADJN (error en el número del adjetivo)*

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

La figura anterior ejemplifica el error GADJN en nueve ocasiones. El error consiste en usar morfemas plurales con el adjetivo.

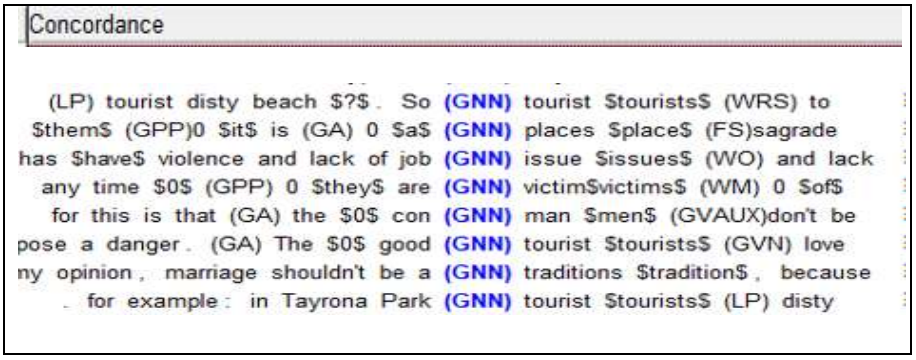


Figura 8. *Alineación de error GNN (error en el número del nombre)*

En el anterior caso, se ejemplifica el error GNN que se refiere a errores en la adición u omisión del morfema plural. Igual que en los casos anteriores la etiqueta de error tiene un contexto que facilita el entendimiento y el análisis de cada caso. Se evidencia en las figuras 6, 7, y 8 que el software de extracción permite la alineación de los errores con un contexto que permite hacer un análisis más acertado de las posibles causas de error si así se desea. Si el investigador desea ver más contexto mientras hace el análisis, simplemente con un clic se puede visualizar el párrafo total y así se logra hacer un análisis más acertado. Esta alineación también permite obtener las estadísticas exactas del número de errores por categoría y por tipo de error. Todos los pasos y procesos descritos en la recopilación de datos siguiendo la metodología de corpus lingüístico y la metodología de análisis de errores en su descripción constituyeron las etapas para el desarrollo de este corpus de aprendientes.

4. Conclusiones

A partir de los resultados obtenidos, puede reafirmarse que la relevancia del análisis de la interlengua de estudiantes colombianos radica en la necesidad de teorizar con el fin de mejorar las prácticas de enseñanza y aprendizaje del inglés como lengua extranjera. El

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

presente trabajo hace un aporte científico a la lingüística aplicada por medio de la metodología de la lingüística de corpus y el análisis de errores asistido por computador (Computer-Aided-Error Analysis - CEA) lo cual permite obtener resultados confiables.

En el caso del presente estudio, el corpus construido se ofrece como una guía para la construcción de nuevos corpus de aprendientes y sirve como herramienta en el seguimiento de los procesos de aprendizaje de lenguas extranjeras. La elaboración de estos corpus junto con los estudios sobre el análisis de errores ayudan a clarificar las etapas que se llevan a cabo en la adquisición de una lengua y permiten entender cómo se internalizan las reglas gramaticales. Los corpus computacionales de aprendientes pueden ser utilizados para hacer análisis semánticos, lexicográficos, pragmáticos, sociolingüísticos, gramaticales, de registro, de distinción de dialectos, de estilo, de estudios literarios, de análisis del discurso, de lingüística forense, etc.

Por su parte, el etiquetador asegura la consistencia en el etiquetado de errores, lo cual garantiza la confiabilidad en los resultados. El uso de etiquetadores de errores estándar permite la comparación de resultados con investigaciones que utilicen la misma metodología, por lo cual resulta necesario hacer más análisis con métodos estandarizados. Cabe aclarar que el uso de un corpus de aprendientes permite buscar patrones de errores. Sin embargo, cuando no se encuentra un patrón, no significa que los alumnos ya hayan adquirido ese tipo de estructura. Por ende, es necesario realizar un análisis más profundo para sacar este tipo de conclusiones.

Referencias bibliográficas

1. Alba Quiñones, V. de. (2009). La enseñanza del español en centros de secundaria alemanes: análisis de errores semánticos. *RedELE*, (16), 1–14. Recuperado de https://www.educacionyfp.gob.es/dam/jcr:38209b23-fdbb-4c2f-83db-f760939b6c30/2009-redele-16-03dealba-pdf.pdf

2. Alexopoulou, A. (2005). Aproximación al tratamiento del error en la clase de E/LE desde la perspectiva del análisis de errores. *Estudios de Lingüística Aplicada*, *23*(41).

3. Bolaños Cuéllar, S. (2015). La lingüística de corpus: perspectivas para la investigación lingüística contemporánea. *Forma y Función, 28*(1), 31–54. https://doi.org/10.15446/fyf.v28n1.51970

4. Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

5. Cabrera, A., Elejalde, J. & Vine, A. (2014). Análisis de errores asistido por computador basado en un Corpus de Aprendientes de Español como Lengua Extranjera. *Revista Signos*, *47*(86), 385–411. https://doi.org/10.4067/S0718-09342014000300003

6. Corder, P. (1981). The Idiosyncratic dialects and error analysis. In *Error analysis and interlanguage*. Oxford: Oxford University Press.

7. Council of Europe (2001). *The common European framework of reference for languages: learning, teaching, assessment.* Cambridge, UK: Press Syndicate of the University of Cambridge.

8. Crespo, M. (2016). Analysis of parameters on author attribution of Spanish electronic short texts. *Research in Corpus Linguistics*, *4*, 25-32.

9. Crystal, D. (2008). *Dictionary of Linguistics and Phonetics* (6th edition). Malden, MA: Blackwell Publishing.

10. Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J. & Thewissen, J. (2005). *Error Tagging Manual Version 1.2.* Centre for English Corpus Linguistics, Université Catholique de Louvain.

12. Garrido, C. (2012). Errors in the Use of English Tenses. *Ikala, Revista de Lenguaje y Cultura, 17*(3), 285–296. https://doi.org/10.1017/CBO9781107415324.004

13. Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & Meunier, F. (Eds.), *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge M.I.T. Press.

14. Granger, S. (2002). A Bird's-eye view of learner corpus research. In Granger, S., Hung, J., & Petch-Tyson, S. (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3-33). Philadelphia: John Benjamins Publishing Company.

15. Granger, S. (2003a). Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO*, *20*(3), 465-480.

16. Granger, S. (2003b). The International Corpus of Learner English : A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly*, *37*(3), 538-546.

17. Hernandez, N. (2013). New media, New Challenges: Exploring the Frontiers of Corpus Linguistics in the Linguistics Curriculum, *Research in Corpus Linguistics, 1*, 17--31.

18. Londoño, D. (2008). Análisis de errores en una composición escrita. *Profile Issues in Teachers' Professional Development, 10,* 135-146.

19. Londoño Vásquez, D. A.. (2008). Error analysis in a written composition. *Profile, 10,* 135-146.

20. López, W. C. (2008). Error Analysis in a Learner Corpus . What Are the Learners' Strategies? *Siglin, 4,* 95-100.

21. McEnery, A, Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book.* London, UK: Routledge.

22. McEnery, Anthony, & Hardie, A. (2011). *Corpus Linguistcis: Method, theory and Practice*. Cambridge: Cambridge University Press.

23. Ortega, M. F. (2014). Assessing Trainees' Oral Performance in a Chilean Teacher Training Program: A Corpus-based Study. *Colombian Applied Linguistics Journal, 16*(1), 10-16. https://doi.org/10.14483/udistrital.jour.calj.2014.1.a01

24. Parada, I., Ruiz, E., & Sánchez, G. (2017). *Prepositional Error Analysis in EFL Students' Written Compositions*. https://doi.org/10.1017/CBO9781107415324.004

25. Saavedra, P., & Campos, M. (2018). Combining the strategies of using focused written corrective feedback: a study with upper-elementary Chilean EFL learners. Combinar las estrategias de uso de retroalimentación correctiva escrita enfocada: un estudio con estudiantes chilenos de EFL de nivel primario superior. *Colombian Applied Linguistics Journal, 20*(1), 79-90. https://doi.org/10.14483/22487085.12332

26. Scott, M. (2008). WordSmith Tools Version 5 [Lexical analysis software]. Liverpool.

27. Sinclair, J. (2005). Developing Linguistic Corpora: a Guide to Good Practice. Retrieved from http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf

28. Szabó, T. (2013). A Corpus-based Analysis of Language Ideologies in Hungarian School Metalanguage. *Research in Corpus Linguistics*, *1*, 65-79.

29. Vázquez, G. (2009). El concepto de error: estado de la cuestión y posibles investigaciones. *Revista Nebrija de Lingüística Aplicada a La Enseñanza de Las Lenguas*, (5). Recuperado de https://www.nebrija.com/revista-linguistica/files/articulosPDF/articulo_53199a284ec8f.pdf

Anexo 1. *Lista completa de errores tomada del manual de errores* (Dagneaux *et al*., 2005).

GDD	Grammar, Determiner, Demonstrative
GDO	Grammar, Determiner, Possessive
GDI	Grammar, Determiner, Indefinite
GDT	Grammar, Determiner, Other
GA	Grammar, Articles
GADJCS	Grammar, Adjectives, Comparative / Superlative
GADJN	Grammar, Adjectives, Number
GADJO	Grammar, Adjectives, Order
GADVO	Grammar, Adverbs, Order
GNC	Grammar, Nouns, Case
GNN	Grammar, Nouns, Number
GPD	Grammar, Pronouns, Demonstrative
GPP	Grammar, Pronoun, Personal
GPO	Grammar, Pronoun, Possessive
GPI	Grammar, Pronoun, Indefinite
GPF	Grammar, Pronoun, Reflexive/Reciprocal
GPR	Grammar, Pronoun, Relative/ Interrogative
GPU	Grammar, Pronoun, Unclear reference
GVAUX	Grammar, Verbs, Auxiliaries
GVM	Grammar, Verbs, Morphology
GVN	Grammar, Verbs, Number
GVNF	Grammar, Verbs, Non-Finite / Finite
GVT	Grammar, Verbs, Tense
GVV	Grammar, Verbs, Voice
GWC	Grammar, Word Class

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54

LCC	Lexis, Conjunctions, Coordinating
LCLC	Lexis, Connectors, Logical, Complex
LCLS	Lexis, Connectors, Logical, Single
LCS	Lexis, Conjunctions, Subordinating
LP	Lexical Phrase
LPF	Lexical Phrase, False friends
LS	Lexical Single
LSF	Lexical Single, False friends

WM	Word Missing
WO	Word Order
WRS	Word Redundant Single
WRM	Word Redudant Multiple

FM	Form, Morphology
FS	Form, Spelling
FSR	Form, Spelling, Regional

QC	Punctuation, Confusion
QL	Punctuation, Lexical
QM	Punctuation, Missing
QR	Punctuation, Redundant

XADJCO	Lexico-Grammar, Adjectives, Complementation
XADJPR	Lexico-Grammar, Adjectives, Dependent Preposition
XCONJCO	Lexico-Grammar, Conjunctions, Complementation
XNCO	Lexico-Grammar, Nouns, Complementation
XNPR	Lexico-Grammar, Nouns, Dependent Preposition
XNUC	Lexico-Grammar, Nouns, Uncountable / Countable
XPRCO	Lexico-Grammar, Prepositions, Complementation
XVCO	Lexico-Grammar, Verbs, Complementation
XVPR	Lexico-Grammar, Verbs, Dependent Preposition

SI	Sentence, Incomplete
SU	Sentence, Unclear

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
E-ISSN 2422-3174
N. º 78, 2020 35-54