SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Detección del nivel de dominio de recursos gramaticales en la redacción de textos técnicos de estudiantes de licenciatura

presentada por

L.I. Leonel González Vidales

como requisito para la obtención del grado de

Maestría en Ciencias de la Computación

Director de tesis

Dr. Noé Alejandro Castro Sánchez

Cuernavaca, Morelos, México. Junio de 2019.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLOGICO NACIONAL DE MEXICO

Centro Nacional de Investigación y Desarrollo Tecnológico

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cuernavaca, Morelos a 21 de junio del 2019
OFICIO No. DCC/064/2019

Asunto: Aceptación de documento de tesis

DR. GERARDO V. GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del Lic. Leonel González Vidales, con número de control M15CE081, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado "Detección del nivel de dominio de recursos gramaticales en la redacción de textos técnicos de estudiantes de licenciatura" y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTOR DE TESIS

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la Computación
08701806

REVISOR 1

Dra. Azucena Montes Rendón
Doctora en Ciencias
4001014

REVISOR 2

Dra. Alicia Martínez Rebollar
Doctora en Informática
7399055

C.p. M.E. Guadalupe Garrido Rivera - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

NACS/lmz

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cuernavaca, Mor., 21 de junio de 2019
OFICIO No. SAC/232/2019

Asunto: Autorización de impresión de tesis

LIC. LEONEL GONZÁLEZ VIDALES
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Detección del nivel de dominio de recursos gramaticales en la redacción de textos técnicos de estudiantes de licenciatura"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

SEP TecNM
CENTRO NACIONAL
DE INVESTIGACIÓN
Y DESARROLLO
TECNOLÓGICO
SUBDIRECCIÓN
ACADÉMICA

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

C.p. Mtra. Guadalupe Garrido Rivera .- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico que me brindó para realizar mis estudios de maestría.

Al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) por brindarme la oportunidad de superarme profesionalmente al formar parte del programa de Maestría en Ciencias de la Computación.

Al Tecnológico Nacional de México y al Instituto Tecnológico de Cd. Altamirano por la beca-comisión que me brindó para realizar mis estudios de maestría.

A los Institutos Tecnológicos de Zacatepec en el estado de Morelos y de Cd. Altamirano en el estado de Guerrero por las facilidades para la realización de las pruebas realizadas en esta investigación.

A mi comité tutorial conformado por la Dra. Azucena Montes Rendón y la Dra. Alicia Martínez Rebollar por sus observaciones y sugerencias para mejorar este trabajo de tesis.

A mi director de tesis, Dr. Noé Alejandro Castro Sánchez, por su consejos y guía para el desarrollo del proyecto.

A la Dra. Leticia Sánchez Lima por su apoyo y sus consejos de redacción de textos técnicos.

¡Muchas gracias a todos!

Dedicatorias

A mi familia:

A mis padres Elida y Leonel, por su apoyo incondicional

en cada uno de los momentos de mi posgrado.

A mis hermanos Jesús y Yajaira por su apoyo y sus consejos.

A mis tías Ma. del Refugio y Angélica por su apoyo.

A mis primas Vielka y Olga por sus consejos.

A mi novia Valeria por iluminar mis días con su amor.

Muchas gracias a todos.

Resumen

En esta investigación se diseñó un algoritmo para identificar el nivel de dominio de los recursos gramaticales en la redacción de textos técnicos de estudiantes de licenciatura.

Este algoritmo consta de las siguientes fases: búsqueda y recuperación de recursos léxicos existentes, desarrollo del módulos de análisis ortográfico, desarrollo del módulo de análisis gramatical y análisis estadístico de la ortografía y gramática.

Adicionalmente, se desarrolló un sistema web para los módulos de análisis ortográfico y gramatical. En este sistema se carga un archivo en formato de texto plano o se escribe un texto técnico directamente en la interfaz. El resultado del análisis fue un reporte divido en tres secciones las cuales contienen: el nivel de dominio de los recursos gramaticales; una estadística de los totales y tipos de errores ortográficos y gramaticales; además de todos los errores ortográficos y los errores gramaticales que se identificaron en el texto técnico analizado.

Los módulos desarrollados se evaluaron haciendo pruebas a 126 documentos obtenidos de estudiantes con licenciatura de I, II, VII y VIII semestres, de los Institutos Tecnológicos de Zacatepec en el estado de Morelos y Cd. Altamirano en el estado de Guerrero. Los resultados que se obtuvieron son los siguientes: la precisión del algoritmo para identificar errores ortográficos y gramaticales es del 95.52 % y su cobertura es del 91.68 %. Además, los resultados indicaron que los errores ortográficos más comunes son de acentuación y los errores gramaticales más comunes son los tipográficos.

Abstract

In this research, an algorithm to identify the level of mastery of grammatical resources in the writing of technical texts of undergraduate students was designed.

This algorithm consists of the following phases: search and retrieval of existing lexical resources, development of the orthographic analysis modules, development of the grammatical analysis module and statistical analysis of spelling and grammar.

Additionally, a web system was developed to perform spelling and grammar analysis. In this system a file in plain text format is loaded or a technical text is written directly in the interface. The result of the analysis was a report divided into three sections which contain: the level of mastery of grammatical resources; a statistic of totals and types of spelling and grammatical errors; in addition to all spelling errors and grammatical errors technical text which was analyzed.

The developed modules were evaluated by testing 126 documents obtained from undergraduate students of I, II, VII and VIII semester, from the Technological Institutes of Zacatepec in the state of Morelos and Cd. Altamirano in the state of Guerrero. The obtained results are the following: the precision of the algorithm to identify spelling and grammatical errors is 95.52 % and your coverage is 91.68 %. Moreover, the results indicated that the most common spelling errors are accentuation errors and the most common grammatical errors are typographic nature.

Índice general

Página

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Los estudiantes que ingresan a cursar una licenciatura requieren tener la competencia de comunicación escrita. Los planes de estudio 2009-2010 por competencias profesionales del Tecnológico Nacional de México (TecNM) tiene una habilidad en el perfil del egresado para el desarrollo de habilidades de comunicación escrita y oral en su propio idioma y en un idioma extranjero. Sin embargo, los planes de estudios no tienen una materia de carácter obligatorio, o por lo menos optativa, que permita desarrollar dicha habilidad. Los profesores suponen que ésta ha sido desarrollada durante el nivel bachillerato, porque en dicho nivel existe la materia de taller de lectura y redacción.

Como resultado de la presente investigación, se diseñó un algoritmo para identificar el nivel de dominio de los recursos gramaticales en la redacción de textos técnicos de estudiantes de licenciatura. El algoritmo recibe textos técnicos de estudiantes de licenciatura que serán analizados por los módulos de procesamiento ortográfico y gramatical, los cuales identificarán errores ortográficos y gramaticales. La salida del procesamiento fue un reporte detallando los errores detectados y la categoría de error a la que pertenecen. El algoritmo que se diseño es de vital importancia porque, mediante el reporte que obtuvo, ofrece información sobre las capacidades de redacción de los estudiantes, con lo cual se podrá identificar su grado de dominio de la gramática castellana con el fin de garantizar que en el futuro inmediato sean capaces de redactar su tesis de grado de manera clara y coherente y con ello mejorar los tiempos requeridos de los escritos para graduarse de los programas licenciatura.

1.1. Descripción del problema

Cuando se redacta un trabajo técnico se requiere de claridad, brevedad y precisión, es decir, que las palabras comuniquen el mensaje de forma clara, fácil de entender y con párrafos bien

construidos. La mala redacción de de un escrito técnico conlleva a cometer algunos de los siguientes errores: errores sintácticos, de concordancia, de puntuación, de redundancia, tipográficos y ortográficos. Todos estos errores pueden impedir o dificultar la conclusión de la tesis y con ello, retrasar la culminación de los estudios.

En esta investigación se desarrolló un algoritmo para identificar el nivel de dominio de los recursos gramaticales en la redacción de textos de estudiantes de licenciatura, lo que permite tener indicios sobre la magnitud de errores y aciertos que los estudiantes llegan a cometer en los escritos que elaboran. Además, se generó un sistema web que analiza los escritos, identifica los errores cometidos, genera estadísticas de éstos y proporciona sugerencias de corrección.

1.2. Objetivos

1.2.1. Objetivo general

Diseñar un algoritmo que permita identificar el nivel de dominio de los recursos gramaticales en la redacción de textos técnicos de estudiantes de licenciatura.

1.2.2. Objetivos específicos

- Diseñar un algoritmo de análisis ortográfico y gramatical de textos escritos.
- Generar un corpus de textos de estudiantes de nivel superior.
- Generar reportes finales de la aplicación del algoritmo a los textos analizados.
- Proponer niveles de redacción para identificar el dominio de características lingüísticas.

Capítulo 2

Marco conceptual

En este capítulo se explicarán los conceptos básicos que sirven de base al desarrollo de este tema de tesis con el propósito de usar como base para el desarrollo del algoritmo para identificar el nivel de dominio de los recursos gramaticales.

2.1. Procesamiento de lenguaje natural

Esta investigación se realizó dentro del campo del Procesamiento de lenguaje natural, el cual Gelbukh y Sidorov (2006, pp. 16) lo conceptualizan como la ciencia que permite a las computadoras comprender los textos por su sentido y no como un archivo binario. Ferreira y Kotz (2010) consideran que el Procesamiento de lenguaje natural incluye la comprensión y la generación del lenguaje. Para lograr estos dos aspectos, consideran que el sistema debe conocer todos los niveles de la lengua, convenciones de discurso y de uso.

2.1.1. Niveles de estudio del Procesamiento de lenguaje natural

De acuerdo al análisis que se quiere realizar, existen varios niveles de análisis de lingüística que se utilizan para el Procesamiento de lenguaje natural.

- **Nivel fonético y fonológico.** En este nivel se estudian los sonidos y su representación abstracta (fonemas).
- **Nivel morfológico.** En este nivel se estudian los mecanismos de formación de la adaptación de los lemas al contexto de uso y de las unidades mínimas de modificación de forma (morfemas).
- **Nivel sintáctico.** En este nivel se estudia la forma en cómo se relacionan los conjuntos de palabras en los subconjuntos de una frase (sintagmas) o en la frase en general.

- **Nivel semántico.** En este nivel se centra en tratar el significado de los términos (relación entre significado y significante).
- **Nivel pragmático.** En este nivel se estudia la interpretación del significado según la situación. Alberich (2007)

2.2. Escritura académica

Se conoce como escritura académica, como todos los documentos producidos dentro de los espacios académicos tales como ensayos, resúmenes, exámenes y evaluaciones. Además de todos los textos producidos por académicos para la difusión del conocimiento científico tal como ponencias, artículos científicos, libros, tesis (Fernández-Fastuca y Bressia, 2009). Este tipo de escritura es la que se analiza en la presente investigación

2.3. Recurso gramatical

Según Echeverría-Arriagada (2016) un recurso gramatical es la combinación de los diferentes elementos de la lengua (adverbios, adjetivos, artículos, determinantes, nombres propios, pronombres, sustantivos, verbos, conjunciones y preposiciones) respetando la gramática de lengua. Es decir, es la utilización correcta del lenguaje, lo cual requiere el respeto a las reglas gramaticales y ortográficas. En esta investigación se identifica nivel de dominio de estos recursos.

2.4. Ortografía

La ortografía fue una de las variables que se evaluaron con la presente investigación. Según la Real-Academia-Española (2010b) consiste en un conjunto de las normas que rigen la correcta escritura de una lengua. Estas reglas son de carácter normativo, es decir, deben cumplirse siempre.

2.4.1. Error ortográfico

El algoritmo desarrollado identifica errores ortográficos, los cuales, según ThambiJose (2014), son errores cognitivos que consiste en el uso de una ortografía desviada porque el escritor no conoce la ortografía correcta de una palabra. Dos características importantes de estos errores son: generalmente dan como resultado una palabra que es fonológicamente idéntica o muy similar

a palabras correctas y los nombre propios, las palabras poco frecuentes y las palabras prestadas de otro idioma son propensas a estos errores.

2.5. Gramática

La gramática fue otra de las variables que se evaluaron con la presente investigación. De acuerdo con la Real-Academia-Española (2010a) es la encargada de estudiar la estructura de las palabras, la organización de ellas dentro de las oraciones y el significado de tal organización. Es decir, es la encargada de regir o guiar el uso de cualquier lengua.

2.5.1. Error gramatical

El algoritmo desarrollado identifica errores gramaticales, para lo cual Nordquist (2018) lo conceptualiza como un término que se utiliza para describir un uso defectuoso, no convencional o controvertido de las reglas gramaticales de una lengua. También se denomina error de utilización.

2.6. Procesamiento de texto

Para realizar el procesamiento de texto se utilizó *Freeling* la cual es una librería de código abierto que proporciona un *front-end* y servicios de Procesamiento de lenguaje natural tales como: lematización, etiquetado de categoría gramatical, fragmentación del texto y reconocimiento de entidades nombradas a desarrolladores de aplicaciones PLN (Atserias *et al.*, 2006).

2.6.1. Análisis morfosintáctico

Con el algoritmo desarrollado, se realiza un análisis morfosintáctico, el cual es utilizado para determinar la forma, clase o categoría gramatical de las palabras de una oración (Gelbukh y Sidorov, 2006). El análisis realizado por *Freeling* se representa a través de las etiquetas EAGLES, las cuales fueron propuestas por el grupo EAGLES (*Expert Advisory Group on Language Engineering Standards*) para determinar la categoría gramatical de las palabras de las lenguas europeas (Cirera, 2012). A continuación se muestra un ejemplo el resultado del análisis morfosintáctico de la oración *Las cajas deben tener un peso máximo de 80 kg.*

Las	cajas	deben	tener	un	peso	máximo	de	80_kg	.
DA0FP0	NCFP000	VMIP3P0	VMN0000	DI0MS0	NCMS000	AQ0MS00	SP	Zu	Fp

En la primera fila se muestran las palabras de la oración. En la segunda fila se muestran las etiquetas que representan la categoría gramatical de cada palabra. Ejemplo: en la Tabla 2.1 se muestra la codificación de la etiqueta *VMIP3P0*.

Tabla 2.1. Codificación de la etiqueta *VMIP3P0*

Posición	**Atributo**	**Valores**
0	Categoría	*V*:verbo
1	Tipo	*M*:principal
2	Modo	*I*:indicativo
3	Tiempo	*P*:presente
4	Persona gramatical	*3*:tercera persona
5	Número	*P*:plural
6	Género	*0*:ninguna

Capítulo 3

Estado del arte

En este capítulo, se presentan los los artículos de investigación que están relacionados con la problemática que aborda el presente estudio. De entre la amplia bibliografía, se seleccionaron los que se consideraron más relevantes.

3.1. Análisis de la competencia lingüístico-discursiva escrita de los alumnos de nuevo ingreso del Grado de Maestro en Educación Primaria

Rico y Dimitrinka (2015) realizaron una investigación para la detección y clasificación de las carencias lingüístico-discursivas de los alumnos que cursaban carreras para obtener una maestría en Educación Primaria. Se utilizó una metodología empírico-analítica con un diseño ex-post-facto de tipo descriptivo, la cual utiliza datos cuantitativos y cualitativos para clasificar y cuantificar las carencias lingüístico-discursiva de los alumnos participantes.

Los datos los recogieron en tres sesiones semanales de sesenta minutos cada una durante cuatro meses de trabajo en la asignatura de Didáctica de la Lengua Española. Al final, los alumnos realizaron una prueba individual con duración de dos horas.

Para el análisis de la información realizaron una revisión bibliográfica, por lo que se acordó la agrupación de los distintos errores encontrados en los textos según: el rendimiento académico de la asignatura, nivel fonológico, nivel morfológico, nivel sintáctico, nivel léxico-semántico y nivel discursivo. Con la ayuda del programa informático *Statistical Package for Social Sciences (SPSS)* se realizó una estadística descriptiva de la muestra participante.

Como resultado, observaron que existen diferencias significativas a favor de las mujeres. Además, se han observado carencias lingüísticas del alumnado que debieron subsanarse en etapas

previas. Tales carencias fueron:

- Las referentes a la segmentación de la palabra en sílabas.
- Diferenciación de categorías gramaticales.
- De unidades sintácticas, o de ortografía.
- Subordinación de oraciones.
- Faltas de ortografía grafofónica.
- Problemas de acentuación y de puntuación.

En la redacción, se encontraron inconvenientes en dotar de coherencia y cohesión a los textos. Esto genera que el alumno emplee una mezcla de lenguaje formal e informal, o de registro oral y escrito.

El análisis de los escritos lo realizaron de forma manual. Esto puede provocar errores debido a algún despiste u omisión.

3.2. Las prácticas de evaluación docente y las habilidades de escritura requeridas en el nivel posgrado

Ramos (2014) realizó una investigación acerca de la importancia de la escritura académica para desarrollar el pensamiento crítico de los estudiantes de posgrado a través de un estudio exploratorio de tipo descriptivo.

Analizaron tres aspectos:

1. La importancia y el peso de la escritura en la evaluación.
2. Las tipologías textuales solicitadas.
3. Los criterios que los docentes emplean para evaluar los trabajos escritos.

La encuesta estuvo divida en dos secciones con 33 descriptores: un cuestionario abierto sobre las prácticas que utilizan los docentes para evaluar el desempeño académico de los estudiantes

y una encuesta tipo *Likert* sobre las habilidades de escritura que los docentes requieren para considerar que sus alumnos sean académicamente competentes. La encuesta se generó en *Google Forms* y se distribuyó por correo electrónico a 800 docentes pertenecientes a 40 programas de posgrado de la Universidad Nacional Autónoma de México (UNAM).

Los resultados que obtuvieron fueron que los docentes del posgrado del área de ciencias físico-matemáticas e ingenierías respondieron que utilizan las siguientes prácticas de evaluación:

- Tareas (21 %)
- Examen escrito (40 %)
- Exposiciones orales (13 %)
- Proyecto o trabajo final (16 %)
- Participación o asistencia (9 %)
- Otras actividades (1 %)

En la formación de la competencia comunicativa los docentes consideran las siguientes habilidades como primordiales: organizar las ideas y la información de manera coherente; evitar errores ortográficos; escribir de manera precisa; presentar los datos con claridad; revisar y editar el texto; escribir con fluidez y utilizar correctamente la gramática del español. Las habilidades consideradas para la construcción del conocimiento son: dar crédito a las fuentes, abstraer información esencial, analizar y sintetizar información de múltiples fuentes e integrar debidamente el material citado y referenciado en el texto.

Este autor concluye que "la competencia comunicativa escrita en español y la construcción del conocimiento son indispensables para el éxito académico de los estudiantes de posgrado. No obstante, se requiere que dichas habilidades sean enseñadas de manera explícita en los diversos ciclos universitarios"pp. 168.

3.3. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods

Tausczik y Pennebaker (2010) realizaron una revisión de cómo fue creado y validado el software Linguistic Inquiry and Word Count (LIWC). LIWC es un programa de análisis de texto que cuen-

ta las palabras en categorías psicológicamente significativas. Los resultados empíricos utilizados en LIWC demuestran su capacidad para detectar el sentido de una amplia variedad de entornos experimentales, tales como: la emotividad, las relaciones sociales, los estilos de pensamiento y las diferencias individuales. LIWC tiene dos elementos centrales: el módulo de procesamiento y los diccionarios. Sigue el siguiente procedimiento para realizar el análisis de textos:

1. Al introducir un texto, el software analiza palabra por palabra y revisa que esas palabras estén en su diccionario.

2. Si la palabra está contenida en el diccionario, la asocia con una clasificación (verbos auxiliares, pronombres personales, etc.).

3. Al final, calcula el porcentaje de cada categoría.

No se presentan resultados, sólo conclusiones de la importancia del análisis de textos.

3.4. La precisión gramatical mediada por la tecnología: el análisis y tratamiento automático de errores

Kotz y Ferreira (2012) realizaron una investigación de la descripción del reconocimiento y tratamiento de errores gramaticales a través de un analizador sintáctico de oraciones en el contexto de un Sistema Tutorial Inteligente (STI) para la enseñanza del español como lengua extranjera. a través de un analizador automático (*parser*).

El objetivo del *parser* es analizar la entrada del usuario y compararla con las gramáticas del sistema. Si es correcto se da un *feedback* positivo. De lo contrario se muestra un *feedback* negativo. Los *feedback* se muestran mediante reportes que dan muestra de los resultados encontrados.

Este artículo no muestra resultados obtenidos, sólo ejemplos del funcionamiento del tutor con frases de ejemplo.

El corpus utilizado tiene limitaciones al tratar de manera automática algunos errores. El *parser* no detecta errores de origen léxico-semántico.

3.5. Un corpus de bigramas utilizado como corrector ortográfico y gramatical destinado a hablantes nativos de español

San Mateo (2016) aplicó el análisis estadístico de la frecuencia de las palabras y de los pares de palabras ('bigrama') utilizados para la detección y corrección de errores ortográficos y gramaticales en textos escritos por nativos. Comparó un corpus textual en español, de cien millones de vocablos, con un texto escrito. Para la detección y corrección de errores ortográficos y gramaticales utilizó una metodología sencilla: comparó las combinaciones de palabras (bigramas) con un corpus de textos y detectaron si estas combinaciones son poco a nada frecuentes. Al hacerlo, demostraron si existe, o no, un error. El algoritmo estimó la probabilidad del par al tener en cuenta la frecuencia de cada una de las dos palabras por separado, en el corpus.

Para verificar su eficacia comparó tres diferentes correctores (Microsoft Word, SpanishChecker y Stilus) en cinco tipos de errores (errores gramaticales u ortográficos, problemas de paronimia, omisión de palabras, confusión de letras y omisión/inclusión de letras). Este algoritmo detectó el error en el 100 % de los casos, en comparación con el promedio de errores detectados por: 25 % de Microsoft Word, 37 % de SpanishChecker y 23 % de Stilus.

Sin embargo, el algoritmo tiene limitaciones al detectar palabras que son poco frecuentes tales como: identificar un sustantivo en singular mientras que el verbo aparece en plural, no identifica errores del uso de tiempos verbales y en las oraciones con sujeto compuesto. Esto genera falsos positivos y falsos negativos.

El autor menciona que este problema puede mejorarse al usar trigramas o n-gramas mayores.

3.6. Automatic syllabification for Spanish using lemmatization and derivation to solve the prefix's prominence issue

Hernandez-Figueroa *et al.* (2013) proponen un algoritmo para dividir una palabra en silabas. El algoritmo implementa las reglas básicas de silabación combinada con información morfológica y léxica obtenida de tres fuentes: un lematizador, una base de datos y el Corpus de Referencia del Español Actual (CREA) de la Real Academia Española. Este algoritmo intenta dar solución al

problema de los prefijos según su prominencia.

Los pasos del algoritmo propuesto son los siguientes: primero se lematiza la palabra al buscarla en una base de datos léxica. Si se localiza esa palabra, el lema es retornado al proceso. Si la palabra no es localizada, se remueve el prefijo y se busca el resultado en la base de datos léxica. Si no se localiza un prefijo en la palabra, se marca como *desconocida*.

La silabificación se desarrolla en tres formas: si el sistema devuelve una palabra desconocida, se realiza sin tener en cuenta los prefijos. Si se devuelve una palabra conocida, se realiza la silabificación al dividir los prefijos. Si la palabra es recién conocida, se realiza la silabificación al analizar las relaciones derivadas en busca de prefijos ocultos.

Realizaron pruebas al usar el Corpus del Español Actual (CREA) el cual contiene 737,799 palabras diferentes. Reconocieron el 356,185 (48.3 % de las palabras) y 381,614 no fueron reconocidas.

3.7. Tabla comparativa del estado del arte

En la Tabla 3.1 se muestra la comparación entre las aportaciones de las investigaciones relacionadas que se consideraron más significativas y se tomaron en cuenta para integrar el estado del arte del presente tema de tesis.

Tabla 3.1. Tabla comparativa del estado del arte

Documento	Idioma	Tipo de análisis efectuado	Software utilizado	Método de solución	Resultados
Análisis de la competencia lingüístico-discursiva escrita de los alumnos de nuevo ingreso del Grado de Maestro en Educación Primaria	Español	Análisis de textos discursivos	SPSS	Metodología empírico-analítica con un diseño ex-post-facto de tipo descriptivo	Se identificaron las siguientes carencias en el alumnado: las referentes a la segmentación de la palabra en sílabas, diferenciación de categorías gramaticales, de unidades sintácticas, o de ortografía, subordinación de oraciones, faltas de ortografía grafofónica y problemas de acentuación y de puntuación.

Sigue en la página siguiente.

Documento	Idioma	Tipo de análisis efectuado	Software utilizado	Método de solución	Resultados
Las prácticas de evaluación docente y las habilidades de escritura requeridas en el nivel posgrado	Español	Habilidades primordiales de un escrito	Encuesta electrónica E-mail	Estudio exploratorio de tipo descriptivo	Habilidades primordiales en la formación de la competencia comunicativa escrita: organizar las ideas y la información de manera coherente, evitar errores ortográficos, escribir de manera precisa, presentar los datos con claridad, revisar y editar el texto, escribir con fluidez y utilizar correctamente la gramática del español.
La precisión gramatical mediada por la tecnología: el análisis y tratamiento automático de errores	Español	Análisis de errores gramaticales.	ELE-Tutor	Analizador automático.	Generador de retroalimentación de positivos y negativos. No se dan resultados de precisión en la detección.
Un corpus de bigramas utilizado como corrector ortográfico y gramatical destinado a hablantes nativos de español	Español	Análisis de errores gramaticales y ortográficos al usar bigramas.	CorrectMe	Uso de bigramas	Precisión del 100 % en la detección 5 tipos de errores (Errores gramaticales u ortográficos, problemas de paronimia, omisión de palabras, confusión de letras y omisión/inclusión de letras). Limitaciones en la detección de palabras poco frecuentes detecta falsos positivos y falsos negativos.
The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods	Inglés	Análisis de textos y de uso de las palabras	LIWC	Módulo de análisis y uso de diccionarios	No presenta resultados, sólo es un análisis de cómo fue creado y de su funcionamiento.
Automatic syllabification for Spanish using lemmatization and derivation to solve the prefix's prominence issue	Español	Palabras	Silabeador TIP	Implementa las reglas básicas de silabación combinada con información morfológica y léxica obtenida de tres fuentes: un lematizador, una base de datos y el Corpus de Referencia del Español Actual (CREA)	Identifica el 48 % del Corpus de Referencia del Español Actual

Sigue en la página siguiente.

Documento	Idioma	Tipo de análisis efectuado	Software utilizado	Método de solución	Resultados
Detección del nivel de dominio de recursos gramaticales en la redacción de textos técnicos de estudiantes de licenciatura	Español	Análisis de errores gramaticales y ortográficos en textos técnicos.	Desarrollo de un prototipo	Se usaron técnicas híbridas (lingüísticas y estadísticas).	Generador de reportes. Precisión del 85.71 % y cobertura de 76.56 % en la detección de errores.

Capítulo 4

Metodología de solución

En este capítulo se presenta la metodología general de solución de esta investigación que consta de cuatro fases: 1) Búsqueda y recuperación de recursos léxicos existentes, 2) Desarrollo del algoritmo para identificar el nivel de dominio de los recursos gramaticales y 3) Pruebas. En la Figura 4.1 se muestra la metodología de solución.

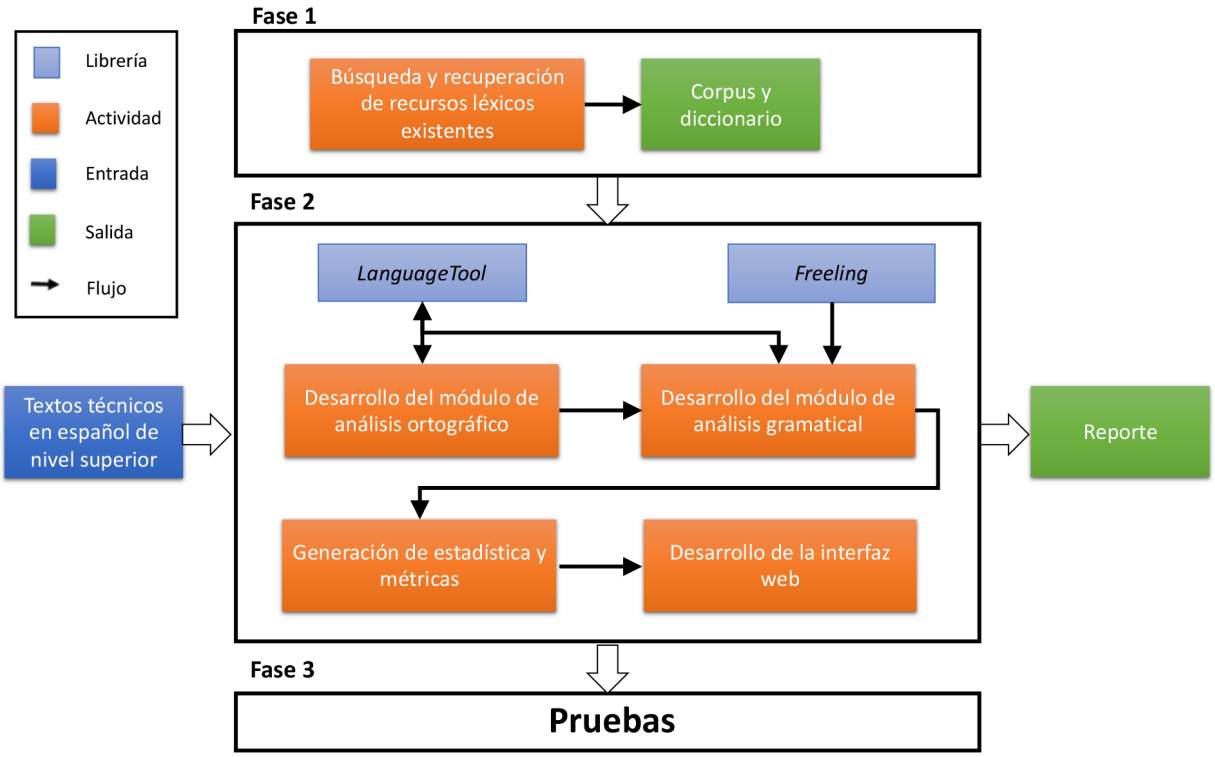

Figura 4.1. Metodología general de solución.

4.1. Fase 1. Búsqueda y recuperación de recursos léxicos existentes

El objetivo de esta fase fue buscar y/o generar recursos léxicos, tales como corpus y diccionarios, para utilizarse como material de apoyo en el desarrollo del algoritmo de análisis ortográfico y gramatical. En la Figura 4.2 se detalla cada una de las etapas.

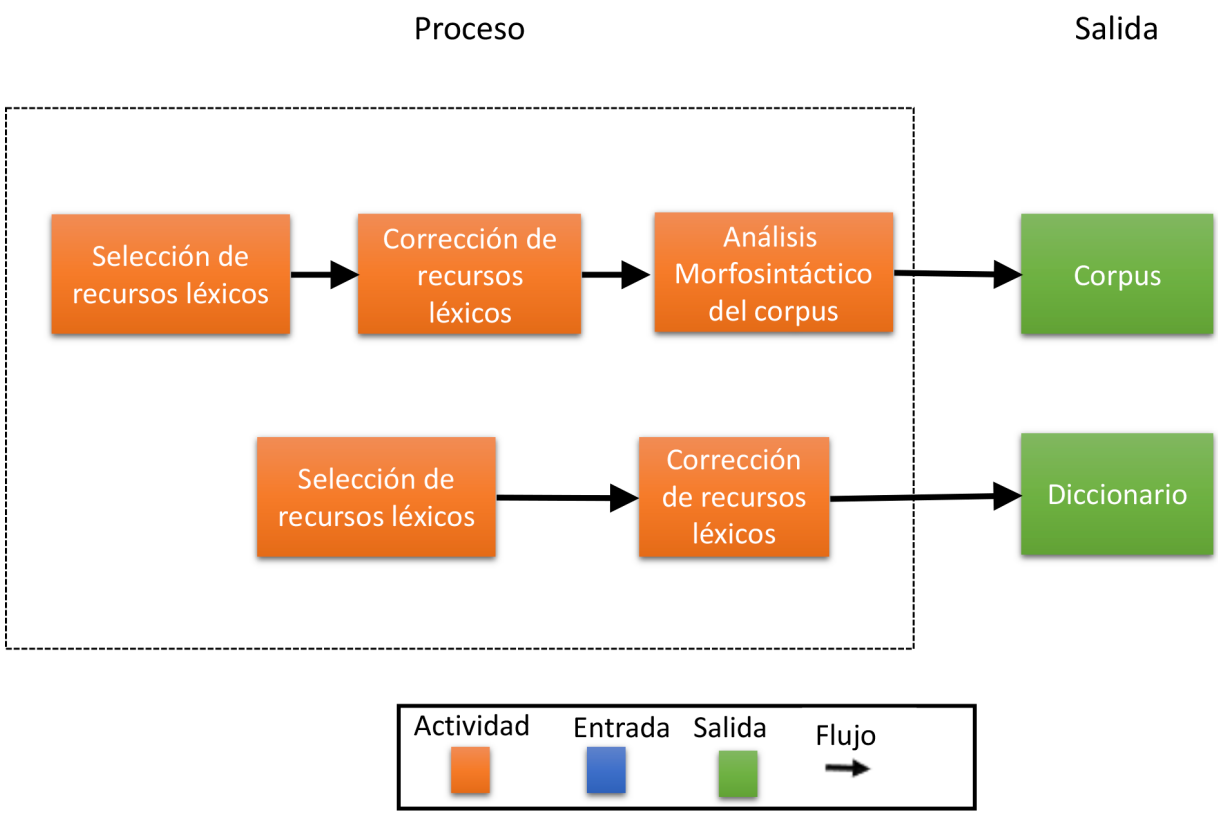

Figura 4.2. Búsqueda y recuperación de recursos léxicos.

4.1.1. Selección de recursos léxicos

En esta fase se realizaron búsquedas en Internet de recursos léxicos en formato de texto plano.

4.1.1.1. Búsqueda de corpus en Internet

Se realizó una búsqueda de corpus textuales de frases en español en formato de texto plano en Internet. En la Tabla 4.1 se muestra la lista de corpus digitales que se visitaron.

Tabla 4.1. Sitios web de corpus en español

No.	Nombre	Página web
1	El corpus del español	(https://www.corpusdelespanol.org/)
2	Corpus del Español: Web/Dialects	https://www.corpusdelespanol.org/web-dial/
3	Real Academia Española - Corpus de Referencia del Español Actual (CREA)	http://corpus.rae.es/creanet.html
4	Corpus del Español Mexicano Contemporáneo (CEMC)	http://www.corpus.unam.mx:8080/unificado/index.jsp?c=cemc
5	European Parliament Proceedings Parallel Corpus 1996-2011	http://www.statmt.org/europarl/

Se seleccionó el corpus paralelo "*European Parliament Proceedings Parallel Corpus 1996-2011*" (Koehn, 2005) porque el corpus está en texto plano. Según estadísticas del sitio, tiene 2,123,835 oraciones y 54,806,927 palabras en español.

4.1.1.2. Búsqueda de diccionarios

Se realizó una búsqueda en Internet del diccionario de Real Academia de Lengua Española en formato TXT. Dado que la Real Academia de la Lengua Española no maneja una presentación en dicho formato, se realizó una búsqueda y se encontró en Domínguez y Valcárcel (2015) tiene el diccionario de la Real Academia Española en su vigésimo tercera edición en formato de texto plano.

El diccionario está dividido en 27 archivos por letra inicial, ejemplo: *a.txt, b.txt, c.txt, ..., z.txt*. Son en total 90,339 palabras.

4.1.2. Corrección de los recursos léxicos

4.1.2.1. Corrección automática del corpus

El corpus tiene 9,433 archivos en formato de texto plano. Los archivos contienen las siguientes etiquetas *XML*: documento (<*CHAPTER id*>), orador (<*SPEAKER id name language*>), y párrafo (<P>). Se realizó una limpieza automática del texto para remover el código *XML*, por medio de expresiones regulares, a cada uno de los archivos en texto plano. En la Figura 4.3

se muestra el resultado de la corrección del corpus. Al terminar, se generó un sólo archivo en formato de texto plano a partir de los 9,433 archivos del corpus.

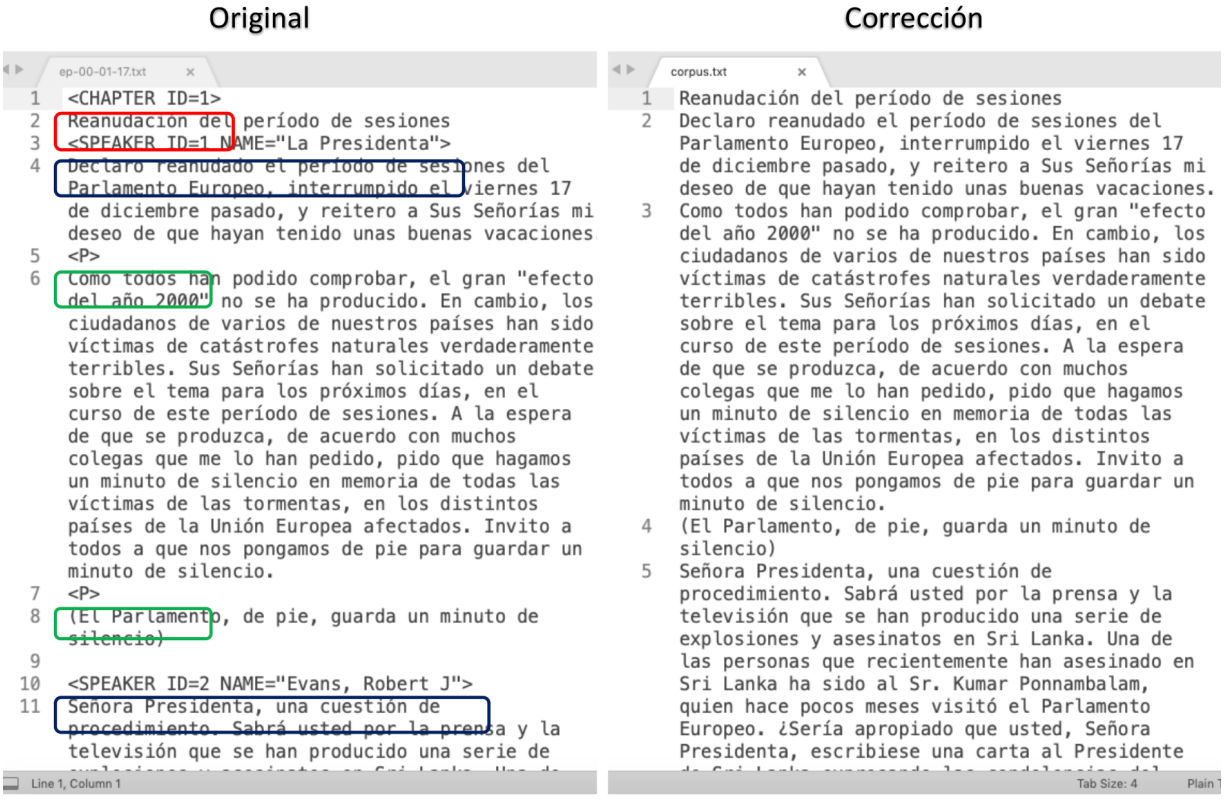

Figura 4.3. Corrección del corpus en formato de texto plano

4.1.2.2. Corrección automática del diccionario

El diccionario de la Real Academia Española en su vigésimo tercera edición en formato de texto plano contenía las palabras en el siguiente formato: *babazorro, rra*, debido a las variantes de genero del español. Por lo que se desarrolló un programa para realizar la adecuación automática de todos los archivos del diccionario. El resultado fue el siguiente: *babazorra, babazorro*. En la Figura 4.4 se muestra el resultado de la adecuación.

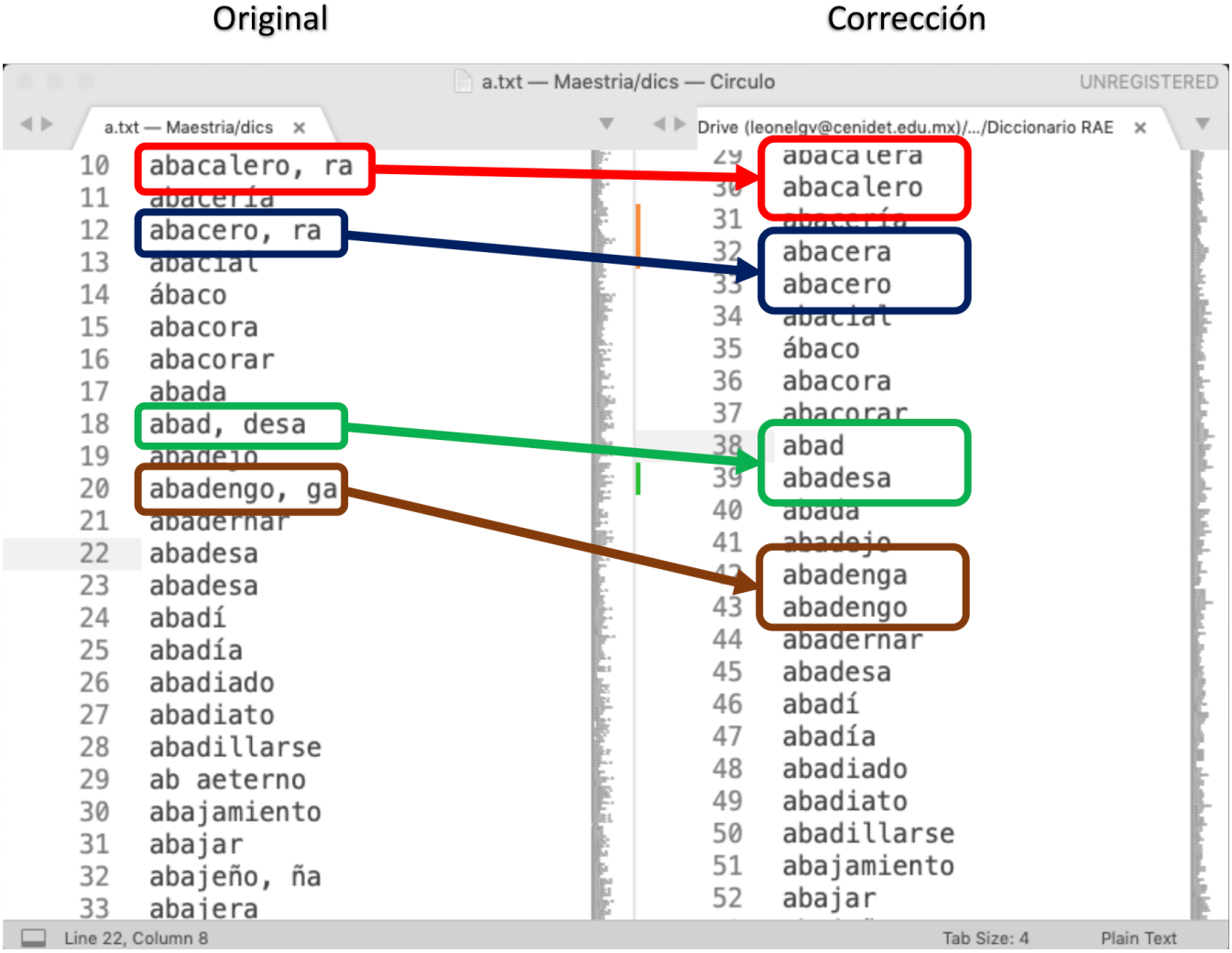

Figura 4.4. Adecuación de los archivos del diccionario en formato de texto plano

4.1.3. Análisis morfosintáctico del corpus

En esta actividad se realizó un análisis morfosintáctico al corpus para determinar la categoría gramatical de cada palabra. Este análisis fue realizado con la librería *Freeling*. Un ejemplo de este análisis se muestra en la Tabla 4.2.

Tabla 4.2. Ejemplo de un análisis morfosintáctico de una oración

Palabra	**Lema**	**Etiqueta**	**Similitud**
La	el	DA0FS0	0.98926
casa	casa	NCFS000	0.998153

Sigue en la página siguiente.

Palabra	**Lema**	**Etiqueta**	**Similitud**
de	de	SP	1
el	el	DA0MS0	1
árbol	árbol	NCMS000	1
es	ser	VSIP3S0	1
pequeña	pequeño	AQ0FS00	0.997312
.	.	Fp	1

El resultado del análisis morfosintáctico del corpus se guardó en un archivo de texto y en una base de datos, la cual contiene 62,475,304 registros. La base de datos se descarga desde la siguiente dirección web: http://bit.ly/2ZHNZ5H. Ésta se utilizó en el desarrollo del algoritmo de análisis gramatical del tema 4.2.2.

4.2. Fase 2. Desarrollo del algoritmo para identificar el nivel de dominio de los recursos gramaticales

El objetivo de esta fase es identificar los errores ortográficos y su tipo de error; los errores gramaticales,su tipo de error. Además de desarrollar el algoritmo para identificar el nivel de dominio de los recursos gramaticales. Las etapas que se llevaron a cabo para el desarrollo del algoritmo para identificar los errores ortográficos y gramaticales se describen con más detalle a continuación.

4.2.1. Desarrollo del módulo de análisis ortográfico

Las etapas que se llevaron a cabo para el desarrollo del módulo de análisis ortográfico se muestra en la Figura 4.5. A continuación se describen con más detalle las actividades que se realizaron en cada etapa.

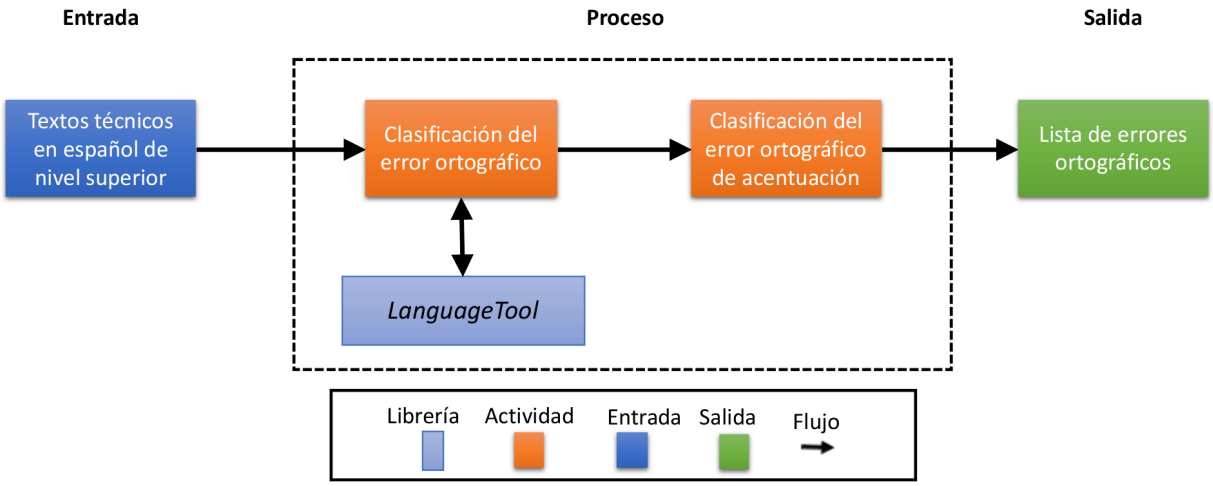

Figura 4.5. Arquitectura del módulo de análisis ortográfico

4.2.1.1. Clasificación del error ortográfico

En esta etapa se identificaron los errores ortográficos. Para lo cual se amplió la funcionalidad de la librería *LanguageTool* (LanguageTool.org, 2016a). La librería genera los siguientes resultados por cada error ortográfico identificado:

- El **párrafo** que contiene el error ortográfico. Ejemplo: "*La utilidad de nuestra carrera es muy amplia puesto que todo lo que nos rodea y lo que somos es practicamente materia viva y día a día se presentan problemas que podemos resolver gracias a la gran amplia capacitación que tenemos a lo largo de nuestra carrera. La carrera de biología le puede dar solución a la gran gama de problemas que existen pensando y preocupandonos siempre por el cuidado, preservación y conservación de la biodiversidad*".
- La **palabra** que es identificada como error ortográfico. Ejemplo: "*practicamente*".
- Un **mensaje** que indica si es un error ortográfico. Ejemplo: "*Error ortográfico*".
- Una **lista de sugerencias** para corregir el error ortográfico. Ejemplo: "*prácticamente, practica mente*".

Además de esto, se agregaron dos funciones más para obtener información adicional del error:

- Una **ventana de contexto** constituida por 2 caracteres a la izquierda y derecha de la palabra identificada como error..., para hacer más fácil su ubicación en el párrafo. Ejemplo: "*somos es* ***practicamente*** *materia viva*".

- La **primera sugerencia** de la lista de sugerencias generada por la librería, pues es la palabra que tiene más probabilidad de ser la correcta. Ejemplo: "*prácticamente*".

La ampliación a la librería consistió generar un análisis que indique cuáles son los tipos de errores ortográficos más comunes en el escrito.

Los errores identificados son:

- **De acentuación**. Uso incorrecto del acento.
- De **sustitución de caracteres por homofonía**. Determina errores en el uso incorrecto por cambios de los siguientes caracteres: *B, V, W, X, S, Z, J, Y, I, Ll, H, R, Rr, Ca, Co, Cu, Ka, Ko, Ku, Ce, Ci, Ze, Zi, Ge, Gi, Je, Ji, Mb, Mp, Nb, Np, Gu, Hu, K, Qu, Gue, Gui.*. Ejemplo: escribir *espectativas* en lugar de *expectativas*.
- De **sustitución de caracteres sin homofonía**. Cambiar una letra por otra. Ejemplo: escribir *canyar* en lugar de *cantar*.
- De **omisión de caracteres**. Identifica errores por omitir o suprimir una letra. Ejemplo: escribir *construción* en lugar de *construcción*.
- De **adición de caracteres**. Identifica errores por agregar letras innecesarias. Ejemplo: escribir *perod* en lugar de *pero*.

4.2.1.2. Clasificación del error ortográfico de acentuación

En esta etapa se realizó una mejora al módulo de análisis ortográfico para identificar el tipo de error de acentuación.

- **De acentuación**. Uso incorrecto del acento.
  - **Palabra aguda no acentuada**. Uso incorrecto del acento en la última sílaba. Ejemplo: escribir *construccion* en lugar de *construcción*.
  - **Palabra grave no acentuada**. Uso incorrecto del acento en la penúltima sílaba. Ejemplo: escribir *seria* en lugar de *sería*.
  - **Palabra esdrújula no acentuada**. Uso incorrecto del acento en la antepenúltima sílaba. Ejemplo: escribir *parrafo* en lugar de *párrafo*.

- **Palabra sobreesdrújula no acentuada**. Uso incorrecto del acento en la antes de la antepenúltima sílaba. Ejemplo: escribir *practicamente* en lugar de *prácticamente*.

4.2.2. Desarrollo del módulo de análisis gramatical

Las etapas que se llevaron a cabo para el desarrollo del módulo de análisis gramatical se muestran en la Figura 4.6. A continuación se describen con más detalle las actividades que se realizaron en cada etapa.

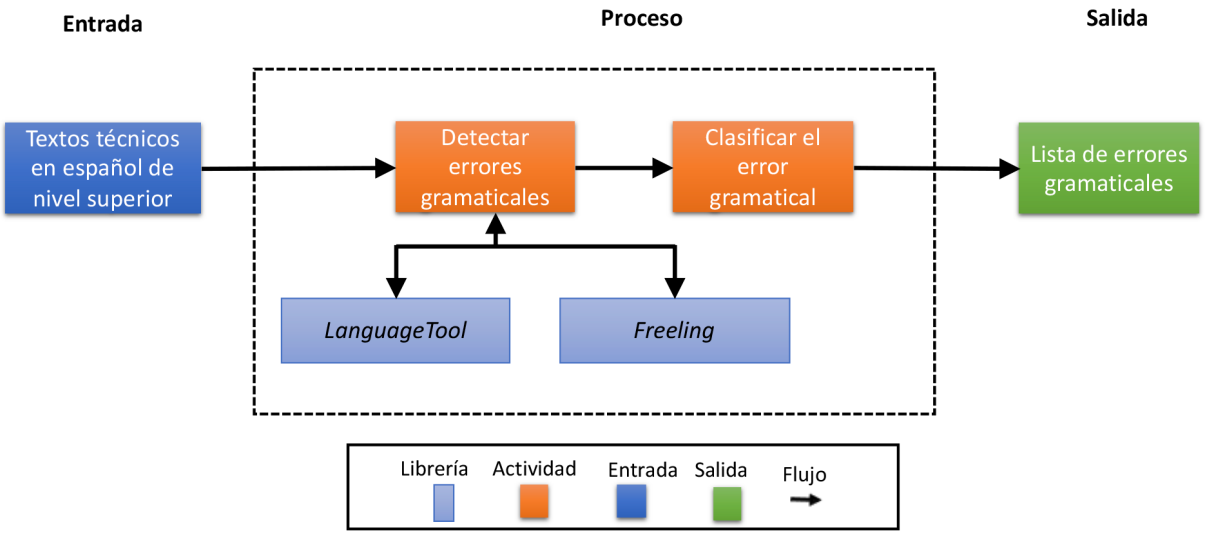

Figura 4.6. Arquitectura del módulo de análisis gramatical

4.2.2.1. Detectar errores gramaticales

En esta etapa se identificaron los errores gramaticales. Para lo cual se amplió la funcionalidad de la librería *LanguageTool* (LanguageTool.org, 2016a). La librería genera los siguientes resultados por cada error ortográfico identificado:

- El **párrafo** que contiene el error gramatical. Ejemplo: "*La utilidad de nuestra carrera es muy amplia puesto que todo lo que nos rodea y lo que somos es practicamente materia viva y día a día se presentan problemas que podemos resolver gracias a la gran amplia capacitación que tenemos a lo largo de nuestra carrera. La carrera de biología le puede dar solución a la gran gama de problemas que existen pensando y preocupandonos siempre por el cuidado, preservación y conservación de la biodiversidad*".
- La **palabra** que es identificada como error gramatical. Ejemplo: "*practicamente*".

- Un **mensaje** que indica el tipo de error gramatical. Ejemplo: "*Error ortográfico*".
- Una **lista de sugerencias** para corregir el error gramatical. Ejemplo: "*prácticamente, practica mente*".

Se agregaron los siguientes resultados para obtener información adicional del error:

- Una **ventana de contexto** constituida por 2 caracteres a la izquierda y derecha de la palabra identificada como error..., para hacer más fácil su ubicación en el párrafo. Ejemplo: "*somos es* ***practicamente*** *materia viva*".
- La **primera sugerencia** de la lista de sugerencias generada por la librería, pues es la palabra que tiene más probabilidad de ser la correcta. Ejemplo: "*prácticamente*".

Adicionalmente se desarrolló un algoritmo para identificar errores gramaticales que no son identificados por la librería. En el apéndice B se muestran aquellos errores gramaticales no identificados. Para esta etapa se realizaron dos pasos:

Mejora del algoritmo de San Mateo (2016)

En el primer paso se realizó una mejora al algoritmo utilizado en San Mateo (2016) para detectar errores gramaticales. En la Tabla 4.3 de se muestra la mejora del algoritmo de San Mateo (2016) la cual se sustituye el análisis de las palabras por su etiqueta morfosintáctica. A continuación se describe el algoritmo utilizado:

Ejemplo:

Tabla 4.3. Ejemplo del algoritmo (San Mateo, 2016) para el bigrama "Sí ,"

No.	**Bigrama**	**Palabra**	**Etiqueta morfosintáctica**
1	*a*	Sí	RG
2	*b*	,	Fc

Se estima la probabilidad del bigrama teniendo en cuenta la frecuencia de cada una de las dos etiquetas, por separado, en el corpus según la fórmula 4.1:

$$P(ab) = \frac{T}{[\frac{T}{f(a)}]x[\frac{T}{f(b)}]} \tag{4.1}$$

- Donde *P(ab)*es la probabilidad del bigrama,
- T es el número total de palabras que componen el corpus (en este caso, 62,475,304).
- *f(a)* es la frecuencia de la etiqueta a (en este caso, 1,904,789).
- *f(b)* es la frecuencia de la etiqueta b (en este caso, 3,113,627).

Ejemplo:

$$P(ab) = \frac{62,475,304}{[\frac{62,475,304}{1,904,789}]x[\frac{62,475,304}{3,113,627}]} = \frac{62,475,304}{32,79906803x20,06512148} = \frac{62,475,304}{658,1172845} = 94,930,3498 \tag{4.2}$$

Tras aplicar la fórmula anterior, se analiza si el bigrama de etiquetas aparece en el corpus más (o menos) veces de lo que sería esperable según su probabilidad –es decir, se calcula el umbral (U) – mediante la fórmula 4.3:

$$U = \frac{F(ab)}{P(ab)} \tag{4.3}$$

- Donde, U es el umbral,
- *F(ab)* es la frecuencia del bigrama (en este caso, 285,735), y
- *P(ab)* es la probabilidad del bigrama (94,930.3498).

Ejemplo:

$$U = \frac{285,735}{94,930,3498} = 3,00994 \tag{4.4}$$

Si el umbral es mayor o igual a uno, las dos etiquetas tienden a usarse juntas. Pero si es menor a uno, las dos etiquetas tienden a rechazarse. Esto se muestra en la Figura 4.7.

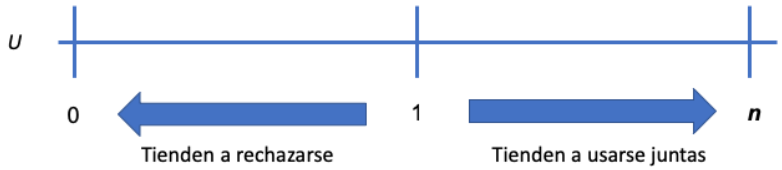

Figura 4.7. Interpretación del umbral

Las palabras seleccionadas se muestran en la Tabla 4.4:

Tabla 4.4. Par de palabras no identificadas correctamente por su contexto

Palabra	**Etiqueta**	**Palabra**	**Etiqueta**	**Ocurrencias**
Si	CS	*Sí*	RG	57
mas	CC	*más*	RG	37
esta	PD0FS00 - DD0FS0	*está*	VMIP3S0	13
practicas	VMIP3S0	*prácticas*	NCFS000	8
ademas	NCFP000	*además*	RG	8
practica	VMIP2S0	*práctica*	NCFP000	8
tenia	NCFS000	*tenía*	VMII3S0	6
como	CS	*cómo*	PT00000	5
maquinas	VMIP2S0	*máquinas*	NCFP000	5
seria	VMIP2S0	*sería*	AQ0FP00	4
calculo	VMIP1S0	*cálculo*	NCMS000	3
computo	VMIP1S0	*cómputo*	NCMS000	3
fabrica	VMIP2S0	*fábrica*	NCFP000	3
publicas	AQ0FS00	*públicas*	VSIC1S0	3
fabricas	VMIP2S0	*fábricas*	NCFP000	1
maquina	VMIP3S0	*máquina*	NCFS000	1

Los resultados de este algoritmo se puede descargar de la siguiente dirección: http://bit.ly/2Ivmfv8.

Algoritmo para identificar errores gramaticales

El segundo paso fue desarrollar el siguiente algoritmo para detectar errores gramaticales. A continuación, se muestra el algoritmo que realiza la identificación de errores gramaticales:

1. El texto se divide en palabras.

2. A cada palabra se le asigna su(s) etiqueta(s) de acuerdo al análisis morfosintáctico (por

ejemplo, coches = sustantivo común masculino plural, hablado = verbo principal participio singular masculino).

3. De las etiquetas asignadas, se eligen las dos primeras categorías (por ejemplo, coches = sustantivo común, hablado = verbo principal).

4. El texto analizado se compara con las reglas incorporadas. Las reglas evalúan el contexto de las palabras (a la izquierda y derecha de la palabra que se evalúa) Lo más importante que debe tener en cuenta es que las reglas describen cómo son las frases correctas.

A continuación se ejemplifica el algoritmo con el texto:

"Desarrollar alumnos competitivos en el ambito laboral.
Si, mayoria de los maestros se encuentra con un nivel de estudios adecuado para impartir las materias, las actividades que se realizan son dinamicas y en muchos casos se llevan a la practica. Y las visitas a empresas donde se menciona cuales son los objetivos de cada una y que actividades realizan para llevarlas a cabo.
Para la solución de problemas de la empresa y para que desarrollarte de una forma competitiva. Todo lo que conlleva una empresa, considerando sus 4 areas RR.HH, Merkadotecnia, Produccion y Finanzas."

A continuación se describen los pasos del algoritmo:

1. Se realiza el análisis morfosintáctico a todo el texto, se extrae la palabra y su etiqueta. Además, se agrega un índice que permite identificar de forma única a cada elemento identificado en el análisis. El resultado se muestra en la Tabla 4.5:

Tabla 4.5. Análisis morfosintáctico del texto

Índice	**Palabra**	**Etiqueta**
0	Desarrollar	VMN0000
1	alumnos	NCMP000
2	competitivos	AQ0MP00

Sigue en la página siguiente.

Índice	Palabra	Etiqueta
3	en	SP
4	el	DA0MS0
5	ambito	NCMS000
6	laboral	AQ0CS00
7	.	Fp
8		
9	Si	CS
10	,	Fc
11	mayoria	NCFS000
12	de	SP
13	los	DA0MP0
14	maestros	NCMP000
15	se	P00CN00
16	encuentra	VMIP3S0
17	con	SP
18	un	DI0MS0
19	nivel	NCMS000
20	de	SP
21	estudios	NCMP000
22	adecuado	VMP00SM
23	para	SP
24	impartir	VMN0000
25	las	DA0FP0
26	materias	NCFP000
27	,	Fc
28	las	DA0FP0
29	actividades	NCFP000

Sigue en la página siguiente.

Índice	Palabra	Etiqueta
30	que	PR0CN00
31	se	P00CN00
32	realizan	VMIP3P0
33	son	VSIP3P0
34	dinamicas	NCFS000
35	y	CC
36	en	SP
37	muchos	DI0MP0
38	casos	NCMP000
39	se	P00CN00
40	llevan	VMIP3P0
41	a	SP
42	la	DA0FS0
43	practica	VMIP3S0
44	.	Fp
45		
46	Y	CC
47	las	DA0FP0
48	visitas	NCFP000
49	a	SP
50	empresas	NCFP000
51	donde	PR00000
52	se	P00CN00
53	menciona	VMIP3S0
54	cuales	PR0CP00
55	son	VSIP3P0
56	los	DA0MP0

Sigue en la página siguiente.

Índice	Palabra	Etiqueta
57	objetivos	NCMP000
58	de	SP
59	cada	DI0CS0
60	una	DI0FS0
61	y	CC
62	que	PR0CN00
63	actividades	NCFP000
64	realizan	VMIP3P0
65	para	SP
66	llevarlas_a_cabo	VMN0000
67	.	Fp
68		
69	Para	SP
70	la	DA0FS0
71	solución	NCFS000
72	de	SP
73	problemas	NCMP000
74	de	SP
75	la	DA0FS0
76	empresa	NCFS000
77	y	CC
78	para	SP
79	que	CS
80	desarrollar	VMN0000
81	te	PP2CS00
82	de	SP
83	una	DI0FS0

Sigue en la página siguiente.

Índice	Palabra	Etiqueta
84	forma	NCFS000
85	competitiva	AQ0FS00
86	.	Fp
87		
88	Todo	PI0MS00
89	lo	DA00S0
90	que	PR0CN00
91	conlleva	VMIP3S0
92	una	DI0FS0
93	empresa	NCFS000
94	,	Fc
95	considerando	VMG0000
96	sus	DP3CPN
97	4	Z
98	areas	NCFP000
99	RR	NP00000
100	.	Fp
101		
102	HH	NP00000
103	,	Fc
104	Merkadotecnia	NP00000
105	,	Fc
106	Produccion	NP00000
107	y	CC
108	Finanzas	NP00000
109	.	Fp
110		

2. Se revisa índice por índice en búsqueda de las siguientes palabras: *practicas, Practicas, practica, Practica, fabrica, Fabrica, fabricas, Fabricas, maquinas, Maquinas, maquina, Maquina, computo, Computo, calculo, Calculo, ademas, Ademas, como, Como, esta, Esta, mas, Mas, publicas, Publicas, publica, Publica, seria, Seria, si, Si, tenia* y *Tenia*.

3. Si se identifica alguna de esas palabras (en este ejemplo, se identificó la palabra *Si* en el índice número nueve) se realizan los siguientes pasos:

    - Se revisa el índice siguiente, en este caso el índice número 10. Si ese índice tiene las etiquetas *AQ, RG, CS, Fit, Fat, Fc, SP, PE, VAI, VAS, VAM, VAC, VAN* o *VAP*, se identifica como error gramatical.
    - Se revisa el índice anterior, en este caso el índice número nueve. Si ese índice tiene las etiquetas *CS, CC, Faa, PP, PD, PI, PR, PE, VSI, VSS, VSM, VSC, VSM, VSP, VSG, VMG* o *vacio*, se identifica como error gramatical.
    - Se extrae el **párrafo** donde aparece el error, la **palabra**, el **contexto de la palabra** y el **tipo de error gramatical**.

4. Si no se encuentra ninguna de la palabras del paso dos, termina.

4.2.2.2. Clasificar el error gramatical

En esta etapa, el módulo de análisis determina el tipo de error gramatical según LanguageTool.org (2016b):

- **Concordancia**. Permite identificar concordancias y discordancias del tipo: de primera, segunda y tercera persona, singular, plural, masculino, femenino. Ejemplos: *el gatos*, *yo hizo*, *cada una de los/las*.
- **Concordancia predictiva**. Permite identificar concordancias de sujeto y predicado en singular, número, femenino y masculino en oraciones atributivas.
- **Diversas**. Permite identificar la repetición de una palabra. Ejemplo: *la casa de de Pedro*.
- **Estilo**. Identifica varias redundancias entre femenino y masculino. Ejemplos: *todas y todos*, *en relación a*, *relacionados a*, *tal es así*.

- **Gramática**. Errores que no cumplen las reglas gramaticales. Ejemplo: intentar combinar una preposición con un verbo conjugado, un no con un imperativo, uso incorrecto del verbo *haber*.
- **Mayúsculas y minúsculas.** Permite identificar cuando la frase se inicia con una letra mayúscula.
- **Ortografía (concepto).** Errores de palabras con sonidos parecidos. Por ejemplo: *haber* por *a ver*, *e* ante palabras empezando por *i*, cambio de *o* ante palabras empezando por *o*.
- **Ortografía (tipográficos).** Errores de palabras que existen en el diccionario, pero se aplican en contextos diferentes. Ejemplo: uso/huso, más/mas, aún/aun, apunto de/a punto de, lo se/lo sé.
- **Posible error tipográfico.**
- **Puntuación.** Identifica dos puntos o comas consecutivos y uso de paréntesis, comillas, signos de exclamación, interrogación y similares disparejos.
- **Tipografía.** Uso de espacios en blanco antes de coma y antes/después de paréntesis.
- **Cambios de normas lingüísticas.** Identifica el uso de sólo/solo y éste/este.

4.2.3. Análisis ortográfico y gramatical

El objetivo de esta fase fue cuantificar la información del texto analizado para disponer de elementos que permitan realizar un análisis. Las etapas que se llevaron a cabo para el análisis ortográfico y gramatical se muestran en la Figura 4.8. A continuación se describen con más detalle las actividades que se realizaron en cada etapa.

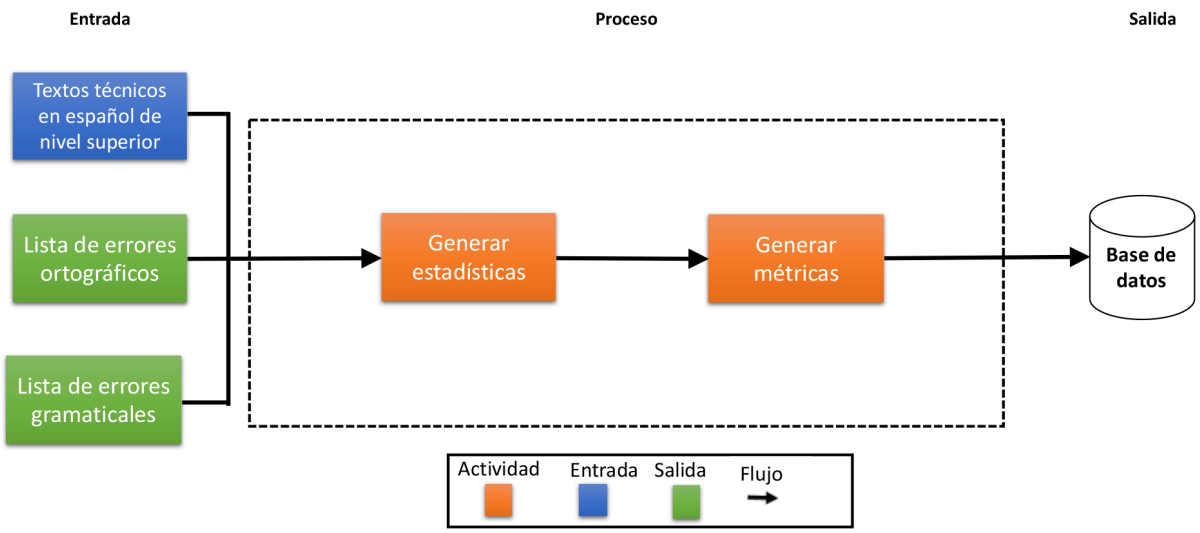

Figura 4.8. Análisis ortográfico y gramatical

4.2.3.1. Generar estadísticas

La herramienta propuesta genera la siguiente estadística.

- El número de párrafos del texto
- El número de oraciones del texto
- La media de oraciones por párrafo
- La desviación estándar de oraciones por párrafo
- El número de palabras del texto
- La media de palabras por oración
- La desviación estándar de palabras por oración

4.2.3.2. Generar métricas

En esta etapa se sugirió una formula para calcular el nivel de dominio de los recursos gramaticales. En la formula 4.5 se calculan el total de errores a través de la suma de los errores ortográficos y gramaticales.

$$totalErrores = totalErroresGramaticales + totalErroresOrtograficos \tag{4.5}$$

Después, se calcula el total de errores de acentuación al agregar un peso mayor a aquellos errores de acentuación que más aparecen en los textos. Los errores de acentuación en la palabras

agudas y esdrújulas tienen un peso de tres. Los errores de acentuación en las palabras graves tienen un peso de dos. Los errores de acentuación en la palabras sobreesdrújulas no tienen peso. El cálculos se muestra en la formula 4.6.

$$\begin{aligned} totalErroresAcentuacion =& (palabrasAgudas * 3) + (palabrasGraves * 2) + \\ & (palabrasEsdrujulas * 3) + (palabrasSobreesdrujulas * 1) \end{aligned} \tag{4.6}$$

Una vez obtenido los dos valores, se suman y se dividen entre el total de palabras del documento (ver la formula 4.7).

$$nivelDominio = \frac{totalErrores + totalErroresAcentuacion}{totalPalabras} \tag{4.7}$$

El resultado obtenido es interpretado de la siguiente forma:

- Si, nivelDominio <0.05, entonces nivel = alto.
- Si, nivelDominio >= 0.05 y nivelDominio <0.1, entonces nivel = medio.
- Si, nivelDominio >= 0.1, entonces nivel = bajo

Por ejemplo:

- Primer semestre:
  - $nivel = \frac{(79+164)+(90*3+64*2+80*3+8*1)}{6199} = 0{,}143410227 \Longrightarrow Nivel\ bajo$
- Último semestre:
  - $nivel = \frac{(93+137)+(102*3+60*2+118*3+9*1)}{8014} = 0{,}127152483 \Longrightarrow Nivel\ bajo$

4.2.4. Desarrollo de la interfaz web

Se desarrolló una interfaz web en donde se implementa el algoritmo para identificar errores ortográficos y gramaticales. Este sistema se nombró ***GrammarChecker***. Para el desarrollo de esta interfaz web se utilizó el lenguaje de programación *Java Server Pages (JSP)*. El sistema interactúa con una base de datos *MySQL* para almacenar los resultados de la evaluación. El sistema web tiene la siguiente dirección de internet: http://tecln.cenidet.edu.mx/grammarchecker/. En la Figura 4.9 se muestra la arquitectura general del sitio web de ***GrammarChecker***.

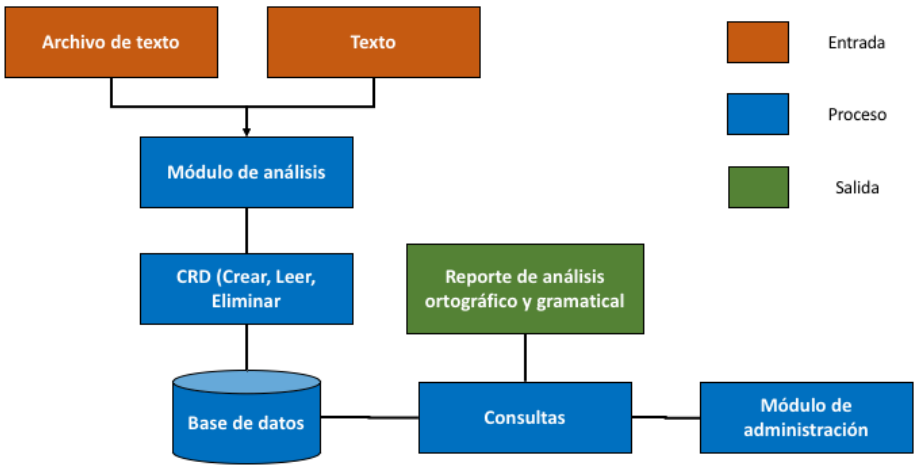

Figura 4.9. Arquitectura del sitio web

4.2.4.1. Módulo de análisis

La interfaz web cuenta con un módulo que permite analizar textos y archivos de texto plano. Desde el módulo se pueden realizar las siguientes operaciones: analizar documento y textos analizados.

En la Figura 4.10 se muestra la pantalla principal del módulo de análisis.

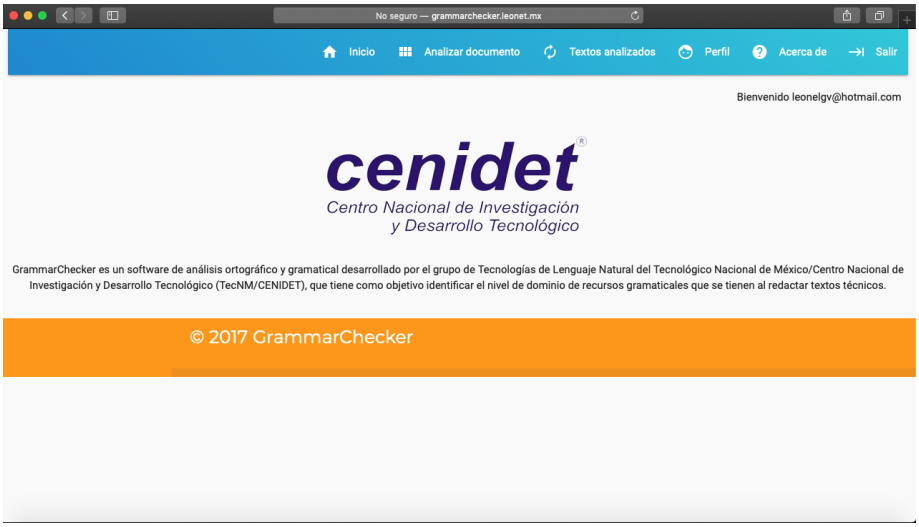

Figura 4.10. Interfaz web principal del módulo de análisis

A continuación se detallan cada una de las funcionalidades del módulo de análisis:

Analizar documento

La interfaz web permite al usuario analizar textos desde dos elementos de entrada: *Escribir texto* y *subir archivo*. A continuación se describe cada uno de los elementos de entrada.

- **Escribir texto**

Desde la interfaz web, el usuario puede seleccionar escribir un texto o subir un archivo de texto plano, posteriormente se analiza el texto y se guardan los resultados en la base de datos. En la Figura 4.11 se muestra la pantalla de la opción *Escribir texto*, en la cual, el usuario escribe un texto en el cuadro de texto, con una longitud máxima de 1500 caracteres.

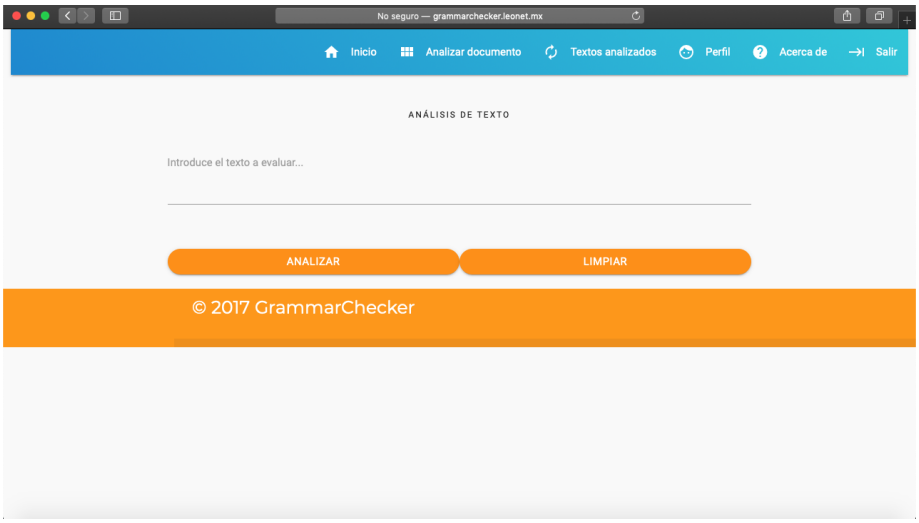

Figura 4.11. Interfaz web: Escribir texto

- **Subir archivo**

En la Figura 4.12 se muestra la opción *Subir Archivo*, en la cual el usuario se le permite subir y analizar un archivo de texto plano. La interfaz permite solamente analizar archivos de textos planos en formato *TXT* que contiene oraciones en español.

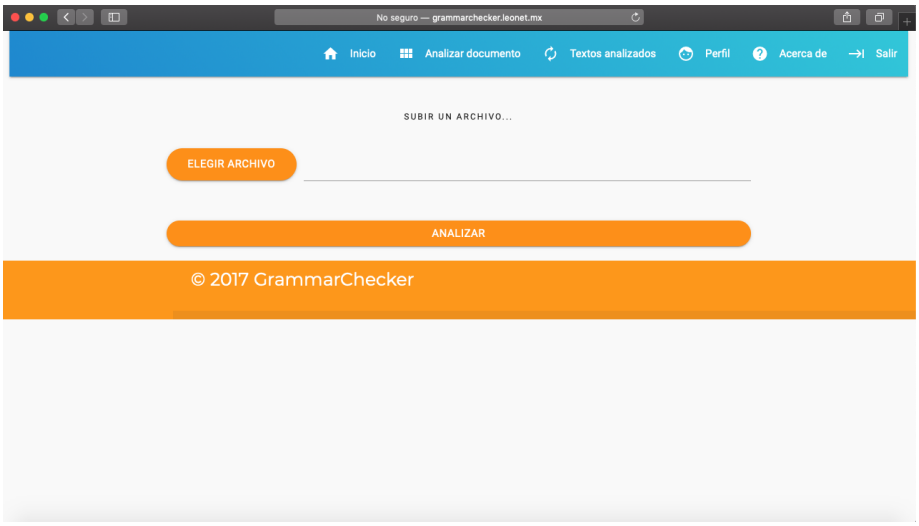

Figura 4.12. Interfaz web: Subir archivo

Textos analizados

La interfaz web muestra sólo los resultados de los textos analizados por el usuario, no podrá ver textos analizados de otros usuarios del sistema web. Además, el usuario podrá descargar los resultados en formato PDF, mostrar los resultados en la interfaz web y borrar el documento analizado. En la Figura 4.13 se muestra la pantalla de *Textos analizados*.

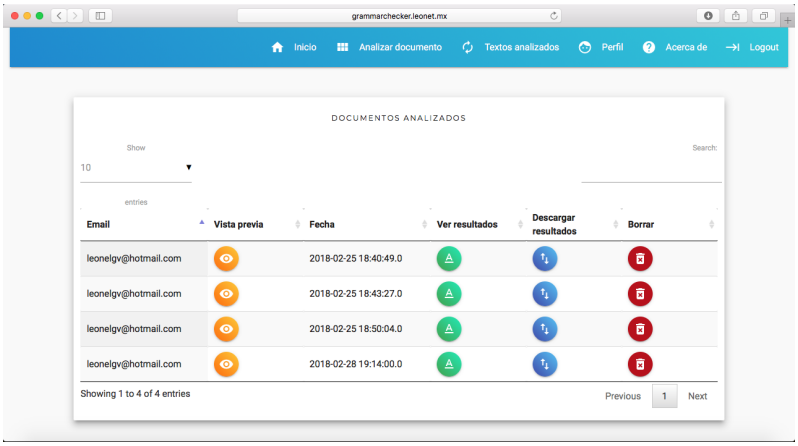

Figura 4.13. Interfaz web: Textos analizados

- **Vista previa**

En esta opción se muestra una vista previa del texto analizado para que el usuario conozca el texto de entrada que utilizó para realizar el análisis. En la Figura 4.14 se muestra la pantalla

vista previa del texto analizado.

Figura 4.14. Textos analizados: vista previa

- **Ver resultados**

En la Figura 4.15 se muestran los resultados en un reporte de análisis ortográfico y gramatical del texto analizado. El reporte está dividido en tres secciones: resultados generales, errores ortográficos y errores gramaticales.

En la primera sección resultados generales se muestra: el nivel de dominio de recursos gramaticales, una estadística básica, los índices de errores ortográficos y gramaticales, el número total de errores ortográficos y gramaticales, el total de errores ortográficos y gramaticales por categoría, el número de errores de sustitución por homofonía y los aciertos en el uso de las categorías de palabras.

Figura 4.15. Reporte de análisis ortográfico y gramatical: resultados generales

En la segunda sección se muestran los errores ortográficos identificados en el texto analizado. Para cada error se muestra el número del error, la palabra con el error ortográfico, una sugerencia que sustituye a la palabra que contiene el error, el tipo de error ortográfico, otras sugerencias para el error ortográfico. Además, se muestra el párrafo que contiene la palabra mal escrita, en el cual, se señala la ubicación exacta dentro del párrafo. En la Figura 4.16 se muestran los errores ortográficos del texto analizado.

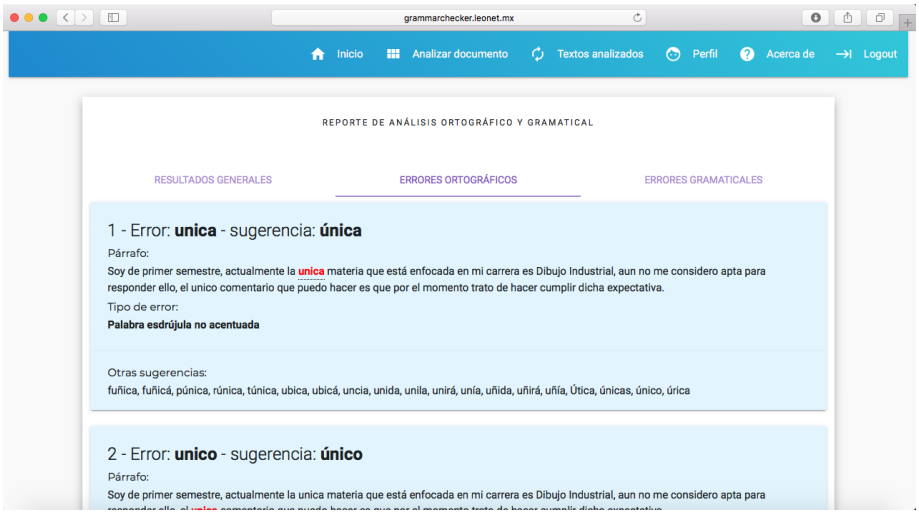

Figura 4.16. Reporte de análisis ortográfico y gramatical: errores ortográficos

En la tercera sección se muestran los errores gramaticales identificados en el texto analizado. Para cada error se muestra el número del error, la palabra con el error gramaticales, una sugerencia que sustituye a la palabra que contiene el error, el tipo de error gramatical, otras sugerencias

para el error gramatical. Además, se muestra el párrafo que contiene la palabra mal escrita, en el cual, se señala la ubicación exacta dentro del párrafo. En la Figura 4.17 se muestran los errores gramaticales del texto analizado.

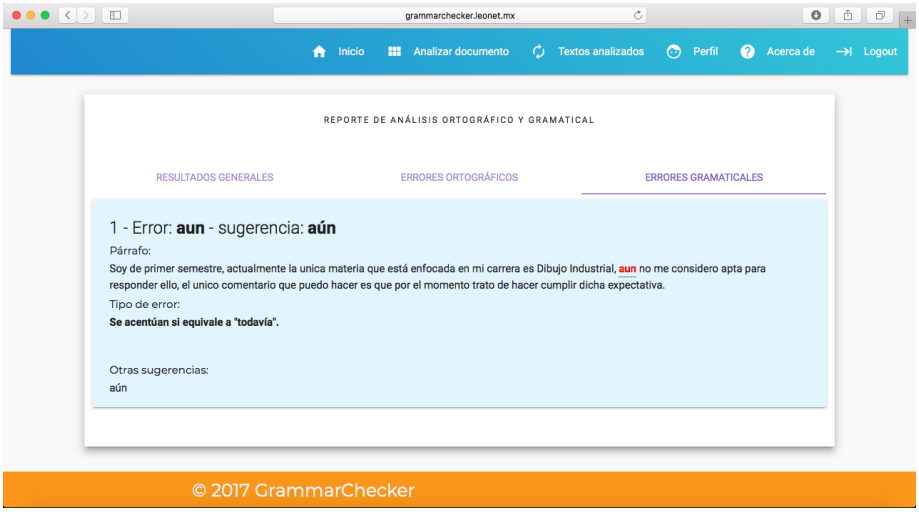

Figura 4.17. Reporte de análisis ortográfico y gramatical: errores gramaticales

- **Descargar los resultados**

En esta opción, el usuario podrá descargar el reporte de análisis ortográfico y gramatical, visto en el tema anterior, en formato *PDF*. Este reporte es de gran ayuda porque permite al usuario imprimirlo para su análisis posterior. En la Figura 4.18 se muestran los resultados en formato *PDF* del texto analizado.

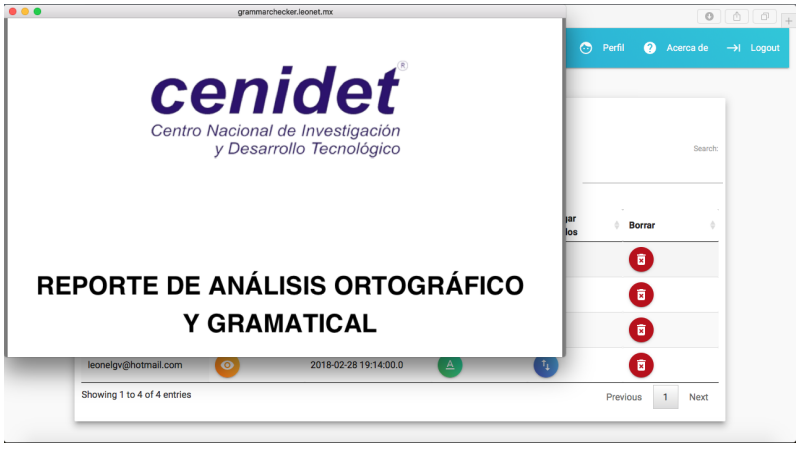

Figura 4.18. Textos analizados: descargar resultados

- **Borrar los resultados**

En esta opción, el usuario podrá borrar el texto analizado y su reporte respectivo para no ser visualizado en futuras ocasiones. En la Figura 4.19 se muestra la opción de borrar el texto analizado y sus resultados.

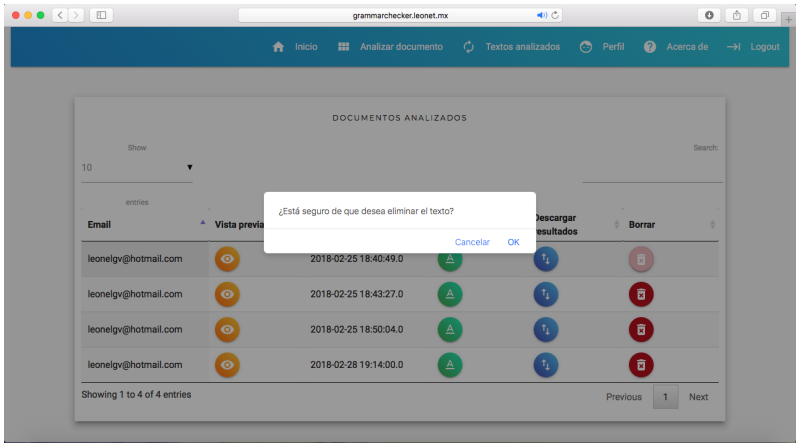

Figura 4.19. Textos analizados: borrar texto analizado

4.2.4.2. Módulo de administración

Se desarrolló un módulo que permite administrar la interfaz web. En éste, el administrador podrá consultar los reportes de análisis ortográfico y gramatical de cada uno de los textos analizados de todos los usuarios del sistema web. Además, contiene una sección para administrar a los usuarios del sistema web.

Textos analizados

En esta sección el administrador visualizará todos los textos analizados de todos los usuarios del sistema. En la Figura 4.20 se muestra un ejemplo de la pantalla principal de la sección en la cual se muestran los resultados de los textos analizados de cada uno de los usuarios del sistema web.

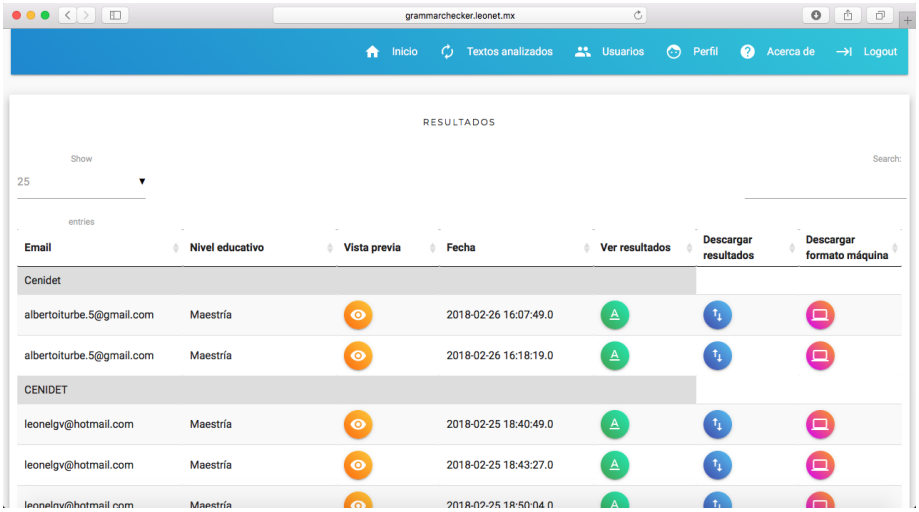

Figura 4.20. Vista administrador. Textos analizados.

A continuación se describen cada una de las opciones de la vista *Textos analizados*.

- **Vista previa**

En esta opción el administrador tiene la opción de ver una vista de cada texto analizado por el sistema. En esta vista sólo se podrá observar el texto capturado o el archivo subido por el usuario. En la Figura 4.21 se muestra la pantalla vista previa del texto analizado.

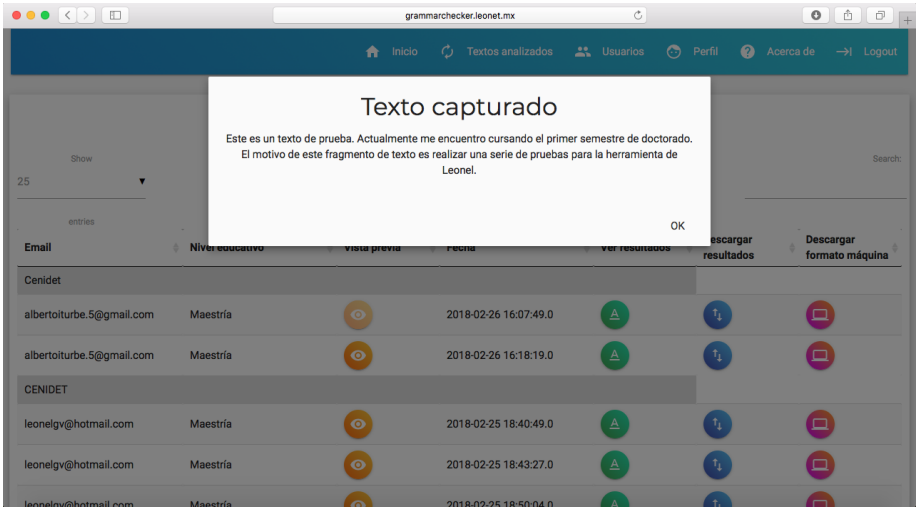

Figura 4.21. Vista administrador. Textos analizados: vista previa

- **Ver resultados**

En esta opción se muestra una ventana con el reporte de análisis ortográfico y gramatical. En la Figura 4.22 se muestran los resultados en un reporte de análisis ortográfico y gramatical del texto

analizado. El reporte está dividido en tres secciones: resultados generales, errores ortográficos y errores gramaticales.

En la primera sección *resultados generales* se muestra: el nivel de dominio de recursos gramaticales, una estadística básica, los índices de errores ortográficos y gramaticales, el número total de errores ortográficos y gramaticales, el total de errores ortográficos y gramaticales por categoría, el número de errores de sustitución por homofonía y los aciertos en el uso de las categorías de palabras.

En la segunda sección se muestran los errores ortográficos identificados en el texto analizado. Para cada error se muestra el número del error, la palabra con el error ortográfico, una sugerencia que sustituye a la palabra que contiene el error, el tipo de error ortográfico, otras sugerencias para el error ortográfico. Además, se muestra el párrafo que contiene la palabra mal escrita, en el cual, se señala la ubicación exacta dentro del párrafo.

En la tercera sección se muestran los errores gramaticales identificados en el texto analizado. Para cada error se muestra el número del error, la palabra con el error gramaticales, una sugerencia que sustituye a la palabra que contiene el error, el tipo de error gramaticales, otras sugerencias para el error gramaticales. Además, se muestra el párrafo que contiene la palabra mal escrita, en el cual, se señala la ubicación exacta dentro del párrafo.

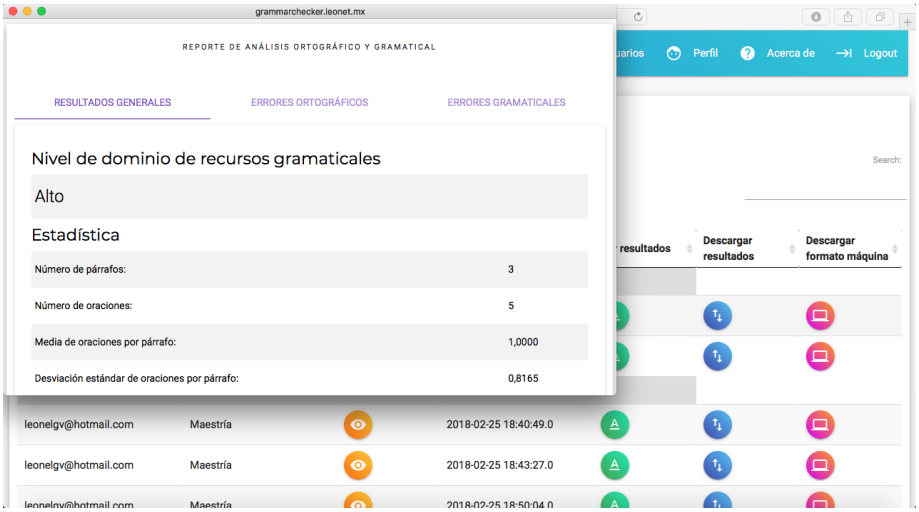

Figura 4.22. Vista administrador. Textos analizados: ver resultados

- **Descargar resultados**

En esta opción, el administrador podrá descargar el reporte de análisis ortográfico y gramatical, visto en el tema anterior, en formato *PDF*. En la Figura 4.23 se muestran los resultados en formato *PDF* del texto analizado.

Figura 4.23. Vista administrador. Textos analizados: descargar resultados PDF

- **Descargar formato legible por computadora**

En esta sección se genera un archivo procesable desde cualquier computadora. Este archivo contiene toda la información mostrada en el reporte de análisis ortográfico y gramatical mostrado en el tema anterior. Además, tiene un formato que puede ser leído por otro sistema o algoritmo para ser utilizado como complemento en su funcionamiento. El formato que se obtiene se muestra en la Figura 4.24:

```
##Número de párrafos##Número de oraciones##Media de oraciones por párrafo##Desviación estándar de oraciones por
párrafo##Desviación estándar de oraciones por párrafo##Número de Palabras##Media de palabras por oración##Desviación
estándar de palabras por oración####Nivel de dominio de recursos gramaticales##Índice error acentuación##Índice error
sustitución sin homofonía##Índice error sustitución por homofonía####Índice error adición##Indice error
omisión##Concordancia##Concordancia predictiva##Diversas##Estilo##Gramática##Mayúsculas y minúsculas##Ortografía (
concepto)##Ortografía (tipográficos)##Posible error tipográfico##Puntuación##Tipografía##Cambios de normas
lingüísticas##Errores detectados##Errores ortográficos##Acentuación de caracteres##Palabras agudas##Palabras
graves##Palabras esdrujulas##Palabras sobreesdrujulas##Sustitución de caracteres sin homofonía####Sustitución de caracteres
por homofonía##Adición de caracteres##Omisión de caracteres####Errores gramaticales####Concordancia##Concordancia
predictiva##Diversas##Estilo##Gramática##Mayúsculas y minúsculas##Ortografía (concepto)##Ortografía (tipográficos)##Posible
error tipográfico##Puntuación##Tipografía##Cambios de normas lingüísticas##Acentuación##Sustitución de B-V##Sustitución de
B-W##Sustitución de V-B##Sustitución de V-W##Sustitución de W-B##Sustitución de W-V##Sustitución de X-S##Sustitución de
S-X##Sustitución de Z-S##Sustitución de S-Z##Sustitución de X-J##Sustitución de J-X##Sustitución de Y-I##Sustitución de
I-Y##Sustitución de LL-Y##Sustitución de Y-LL##Sustitución de H##Sustitución de R-RR##Sustitución de RR-R##Sustitución de
CA_CO_CU-KA_KO_KU##Sustitución de KA_KO_KU-CA_CO_CU##Sustitución de CE_CI-ZE_ZI##Sustitución de ZE_ZI-CE_CI##Sustitución de
GE_GI-JE_JI##Sustitución de JE_JI-GE_GI##Sustitución de MP_MB-NP_NB##Sustitución de NP_NB-MP_MB##Sustitución de
GU-HU##Sustitución de HU-GU##Sustitución de K-QU##Sustitución de QU-K##Sustitución de Ge_Gi-Gue_Gui##Sustitución de
Gue_Gui-Ge_Gi##Aciertos en la acentuación##totales##Aciertos en el uso de B##totales##Aciertos en el uso de
V##totales##Aciertos en el uso de W##totales##Aciertos en el uso de X##totales##Aciertos en el uso de S##totales##Aciertos
en el uso de Z##totales##Aciertos en el uso de J##totales##Aciertos en el uso de Y##totales##Aciertos en el uso de
I##totales##Aciertos en el uso de Ll##totales##Aciertos en el uso de H##totales##Aciertos en el uso de R##totales##Aciertos
en el uso de RR##totales##Aciertos en el uso de Ca-Co-Cu##totales##Aciertos en el uso de Ka-Ko-Ku##totales##Aciertos en el
uso de Ce-Ci##totales##Aciertos en el uso de Ze-Zi##totales##Aciertos en el uso de Ge-Gi##totales##Aciertos en el uso de
Je-Ji##totales##Aciertos en el uso de Mb-Mp##totales##Aciertos en el uso de Nb-Np##totales##Aciertos en el uso de
Gu##totales##Aciertos en el uso de Hu##totales##Aciertos en el uso de K##totales##Aciertos en el uso de
Qu##totales##Aciertos en el uso de Gue-Gui##totales
```

Figura 4.24. Formato legible por computadora generado por el sistema

En la Figura 4.25 se muestra el formato legible por computadora generado por el sistema.

Figura 4.25. Vista administrador. Textos analizados: descargar formato máquina

Usuarios

En esta sección se muestran todos los usuarios registrados en el sistema. Esta vista contiene todos los datos de los usuarios del sistema web. En la Figura 4.26 se muestra la pantalla principal de la administración de los usuarios del sistema web.

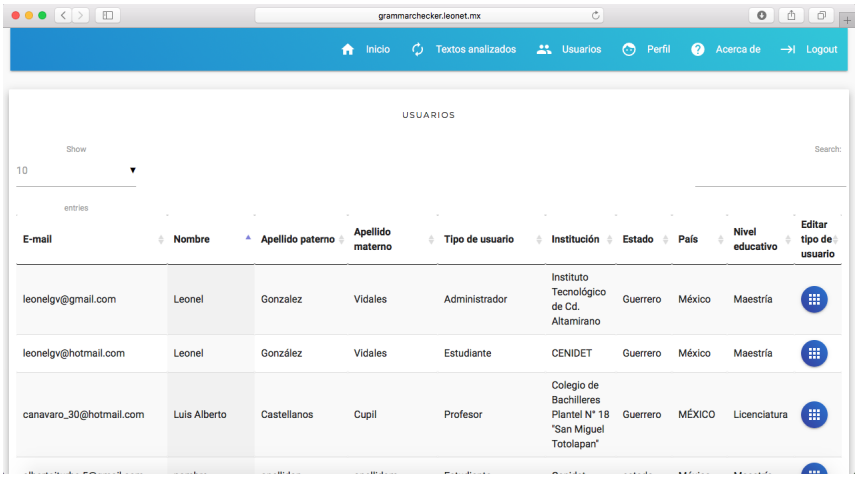

Figura 4.26. Vista administrador: usuarios

- **Editar tipo de usuario**

En esta opción, el administrador del sistema podrá cambiar el tipo de usuario a cualquier usuario del sistema. En la Figura 4.27 se muestra la pantalla principal de la administración del tipo de usuario del sistema web.

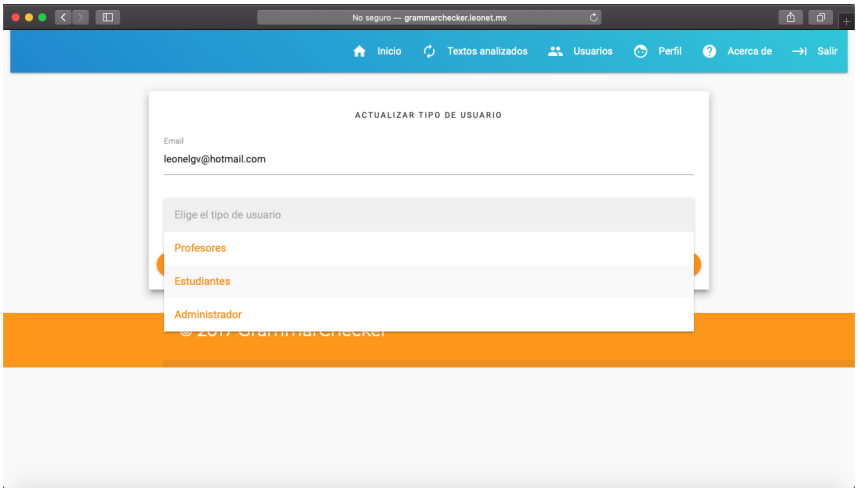

Figura 4.27. Vista administrador: editar tipo de usuario

4.3. Fase 3. Pruebas

En esta fase se realizaron pruebas del funcionamiento del algoritmo. En el siguiente capítulo se explica a detalle.

Capítulo 5

Pruebas y resultados

En este capítulo se presentan las pruebas que se realizaron en esta investigación para obtener resultados como precisión y cobertura.

5.1. Pruebas

Se realizaron pruebas a 126 documentos obtenidos tanto en línea como en forma presencial, con la aplicación de un cuestionario de cuatro preguntas acerca de la satisfacción de los estudiantes sobre su carrera, en los Institutos Tecnológicos de Zacatepec del estado de Morelos y de Cd. Altamirano en el estado de Guerrero (en el Apéndice C se muestra el cuestionario). Los textos fueron revisados manualmente por un experto en gramática y ortografía para identificar los errores y compararlos con los identificados por el algoritmo. Esta revisión se realizó de forma minuciosa para identificarlos.

Los documentos se clasificaron según la Tabla 5.1.

Tabla 5.1. Clasificación de los documentos

Prefijo	**Semestre**	**Ejemplo**	**Total de documentos**
A	Primer semestre	**A**001	63
C	Séptimo semestre	**C**001	38
D	Octavo semestre	**D**001	25

La primera y segunda columna de la Tabla 5.1 representa un código que permite identificar a qué semestre pertenece el documento. En la tercera columna se muestra un ejemplo de cada clasificación. La cuarta columna muestra el total de documentos por cada categoría.

5.2. Resultados

La evaluación de un algoritmo es fundamental no solo para medir su funcionamiento sino para mejorarlo, compararlo o incluso para complementarlo o sustituirlo por otro algoritmo. Para una correcta evaluación del algoritmo que se desarrolló, se cuentan con dos clases de términos: errores y aciertos, con los que se comparan los resultados del algoritmo y se clasifican en distintos grupos:

- verdaderos positivos (VP): errores ortográficos o gramaticales correctamente reconocidos por el algoritmo.
- falsos negativos (FN): errores ortográficos o gramaticales que realmente lo son, pero que el algoritmo indica que no lo son.
- falsos positivos (FP): palabras que cumplen con las reglas de ortografía y gramática, pero que el algoritmo indica que son errores ortográficos o gramaticales.
- verdaderos negativos (VN): palabras correctamente escritas.

En la Tabla 5.2 se muestra la matriz que resume toda esta información: sus filas contienen las clases correctas y sus columnas las clases identificadas del algoritmo.

Tabla 5.2. Matriz para el algoritmo de identificación de errores ortográficos y gramaticales

	Clase identificada	
Clase real	**Errores**	**Aciertos**
Errores	verdaderos positivos *(VP)*	falsos negativos *(FN)*
Aciertos	falsos positivos *(FP)*	verdaderos negativos *(FN)*

La cobertura y precisión son las medidas habituales usadas para evaluar un algoritmo. La cobertura mide la proporción errores correctamente identificados respecto al total de errores reales, dicho de otro modo, mide en que grado *están todos los que son.*

$$cobertura = \frac{VP}{VP + FN} \tag{5.1}$$

La precisión mide el número de errores correctamente identificados respecto al total de los errores identificados, sean verdaderos o falsos errores, dicho de otro modo, mide en que grado *son todos los que están.* (Alcina *et al.*, 2009)

$$precision = \frac{VP}{VP + FP} \tag{5.2}$$

A continuación, se muestran los resultados obtenidos de cada algoritmo.

5.2.1. Resultados de la fase 2

5.2.1.1. Módulo de análisis ortográfico

Se revisaron los 126 documentos con el módulo de análisis ortográfico. Se identificaron un total de 701 errores ortográficos, dos palabras escritas correctamente fueron identificadas como errores y ocho errores ortográficos no se identificaron. En la Tabla 5.3 se muestran los errores ortográficos identificados correctamente por la librería (verdaderos positivos), los errores identificados incorrectamente (falsos positivos) y aquellos errores no identificados (falsos negativos).

Tabla 5.3. Errores ortográficos identificados por el módulo de análisis ortográfico

Categoría	**Totales algoritmo**	**Totales experto**
Errores ortográficos identificados	701	709
Errores ortográficos identificados incorrectamente	2	0
Errores ortográficos no identificados	8	0
Totales	711	709

Con los datos obtenidos, se calculó la medida de precisión y cobertura, tal como se muestra en la Tabla 5.4.

Tabla 5.4. Precisión y cobertura del módulo para detectar errores ortográficos

Tipo de error	**Precisión**	**Cobertura**
Errores ortográficos	99.72 %	98.87 %

Resultados de la clasificación del error ortográfico

Los resultados obtenidos al clasificar los errores ortográficos, el cual utiliza la clasificación vista en el tema 4.2.1.2, fueron los siguientes: se clasificaron correctamente 659 palabras y no clasificó 44 palabras. Los resultados obtenidos por el algoritmo de clasificación de errores ortográficos se muestra en la Tabla 5.5.

Tabla 5.5. Resultados obtenidos del algoritmo de clasificación de errores ortográficos

Categoría	**Totales algoritmo**	**Totales experto**
Errores ortográficos clasificados	659	709
Errores ortográficos clasificados incorrectamente	44	0
Errores ortográficos no clasificados	8	0
Totales	711	709

Se calculó la medida de precisión para medir el número de errores correctamente reconocidos respecto al total de errores predichos y la cobertura mide la proporción de errores correctamente reconocidos respecto al total de errores reales, tal como se muestra en la Tabla 5.6.

Tabla 5.6. Precisión y cobertura del algoritmo de clasificación de errores ortográficos

Librería/Algoritmo	**Precisión**	**Cobertura**
Algoritmo de clasificación de errores ortográficos	93.74 %	98.87 %

Resultados de la clasificación de palabras según su acento

Al clasificar palabras según su acento (agudas, graves, esdrújulas y sobreesdrújulas), se utilizó una lista de 531 palabras acentuadas. El sistema clasificó correctamente las 531 palabras (verdaderos positivos). Los resultados obtenidos de clasificación de palabras según su acento se muestra en la Tabla 5.7.

Tabla 5.7. Resultados obtenidos por el algoritmo de clasificación de palabras según su acento

Categoría	**Totales algoritmo**	**Totales experto**
Errores ortográficos clasificados	531	531
Errores ortográficos clasificados incorrectamente	0	0
Errores ortográficos no clasificados	0	0
Totales	531	531

Se calculó la medida de precisión y cobertura, tal como se muestra en la Tabla 5.8.

Tabla 5.8. Precisión y cobertura del algoritmo de clasificación de errores ortográficos

Librería/Algoritmo	**Precisión**	**Cobertura**
Algoritmo de clasificación de palabras según su acento	100 %	100 %

5.2.1.2. Módulo de análisis gramatical

Se revisaron los 126 documentos con el módulo de análisis gramatical. Se identificaron un total de 103 errores gramaticales, 10 palabras escritas correctamente fueron identificadas como errores y 239 errores gramaticales no se identificaron. En la Tabla 5.9 se muestran los errores gramaticales identificados correctamente por el módulo (verdaderos positivos), los errores identificados incorrectamente (falsos positivos) y aquellos errores no identificados (falsos negativos).

Tabla 5.9. Errores gramaticales identificados por el módulo de análisis gramatical

Categoría	**Totales algoritmo**	**Totales experto**
Errores gramaticales identificados	264	347
Errores gramaticales identificados incorrectamente	43	0
Errores gramaticales no identificados	83	0
Totales	390	347

Se calculó la medida de precisión y cobertura, tal como se muestra en la Tabla 5.10.

Tabla 5.10. Precisión y cobertura del algoritmo para detectar errores gramaticales + *LanguageTool*

	Precisión	**Cobertura**
Algoritmo para detectar errores gramaticales + *LanguageTool*	85.99 %	76.08 %

5.2.1.3. Resultados globales de la fase 2

Los resultados globales fueron los siguientes: se identificaron en total 962 errores ortográficos y gramaticales, 45 palabras escritas correctamente fueron identificadas como error ortográfico o gramatical y 94 errores no se identificaron. Los resultados obtenidos por *LanguageTool* y el algoritmo de detección de errores gramaticales se muestra en la Tabla 5.11.

Tabla 5.11. Resultados obtenidos de los módulos de la fase 2

Categoría	**Totales algoritmo**	**Totales experto**
Errores identificados	965	1056
Errores identificados incorrectamente	45	0
Errores no identificados	91	0
Totales	1,101	1056

Se calculó la medida de precisión y cobertura, tal como se muestra en la Tabla 5.12.

Tabla 5.12. Precisión y cobertura del algoritmo para detectar errores ortográficos y gramaticales + la librería *LanguageTool*

	Precisión	**Cobertura**
LanguageTool + Algoritmo	95.54 %	91.38 %

5.2.2. Nivel de dominio de los recursos gramaticales

En la Tabla 5.13 se muestran los resultados globales de los niveles de dominio de los recursos gramaticales de los estudiantes de licenciatura del primer, séptimo y octavo semestre.

Tabla 5.13. Niveles de dominio de los recursos gramaticales de los estudiantes de licenciatura

Semestre	**Total de estudiantes**	**Nivel bajo**	**Nivel medio**	**Nivel alto**
Primer semestre	63	38	13	12
Último semestre	63	35	19	9
Totales	126	73	32	21

5.2.3. Errores ortográficos y gramaticales más comunes

A continuación se muestran los errores ortográficos y gramaticales más comunes en los que inciden los estudiantes de licenciatura.

5.2.3.1. Errores ortográficos

Los errores ortográficos que más inciden los estudiantes de licenciatura son los de acentuación. En la Tabla 5.14 se muestran los errores ortográficos más comunes.

Tabla 5.14. Errores ortográficos más comunes

Categoría	**Total**
De acentuación	531
Sustitución de caracteres sin homofonía	39
Sustitución de caracteres por homofonía	30
Omisión de caracteres	33
Adición de caracteres	83
Totales	716

De los errores de acentuación, los estudiantes cometen más errores al momento de colocar el acento ortográfico en las palabras esdrújulas y agudas. Sin embargo, se equivocan más al escribir las palabras sobreesdrújulas en un 100 % de los casos, en comparación del 72 % de la palabras esdrújulas y el 33 % de las palabras agudas. En la Tabla 5.15 se muestran los errores ortográficos de acentuación más comunes.

Tabla 5.15. Errores ortográficos de acentuación más comunes

Categoría	**Total palabras mal acentuadas**	**Total palabras bien acentuadas**	**Porcentaje de palabras bien acentuadas**	**Porcentaje de palabras mal acentuadas**
Agudas	192	384	66.67 %	33.33 %
Graves	124	117	48.55 %	51.45 %
Esdrújulas	198	77	28 %	72 %
Sobreesdrújulas	17	0	0 %	100 %
Totales	531	578		

5.2.3.2. Errores gramaticales

Los errores gramaticales que más inciden los estudiantes de licenciatura son los tipográficos. Es decir, aquellos errores de palabras que existen en el diccionario, pero se aplican en contextos diferentes. Ejemplos: *uso/huso*, *más/mas*, *aún/aun*, *apunto de/a punto de*, *lo se/lo sé*. En la Tabla 5.16 se muestran los errores gramaticales más comunes.

Tabla 5.16. Errores gramaticales más comunes

Categoría	**Total**
Concordancia	14
Estilo	4
Gramática	18
Mayúsculas y minúsculas	40
Ortografía (Concepto)	53
Ortografía (tipográficos)	164
Puntuación	3
Tipografía	6
Cambio de normas	3
Diversas	2
Totales	307

Capítulo 6

Conclusiones

Con esta investigación se desarrolló un algoritmo capaz de determinar los tipos de errores ortográficos (de acentuación, sustitución de caracteres por homofonía, sustitución de caracteres sin homofonía, omisión de caracteres y adición de caracteres) que presentan los escritos de los estudiantes de nivel licenciatura y capaz de identificar errores gramaticales que no son identificados por la librería *LanguageTool*. Para aumentar la cobertura en la detección de errores gramaticales se combinó la funcionalidad de la librería *LanguageTool* con el algoritmo desarrollado en este trabajo. Los resultados en la detección de errores gramaticales de la librería *LanguageTool* en comparación con el algoritmo más la librería *LanguageTool* se muestran en la Tabla 6.1.

Tabla 6.1. Comparación de la precisión y cobertura de la librería y de la librería + algoritmo en la detección de errores gramaticales

	Precisión	**Cobertura**
Librería	91.15 %	30.12 %
Librería + Algoritmo desarrollado	85.99 %	76.08 %

En la investigación se identificó que los errores de acentuación son los errores más comunes que cometen los estudiantes de nivel licenciatura, en concordancia con los resultados de Vernon y Alvarado (2013) y a diferencia de lo encontrado en San Mateo (2016) . Por este motivo, el trabajo aquí desarrollado se extendió en la creación de un algoritmo que permite clasificar las palabras agudas, graves, esdrújulas y sobreesdrújulas, lo que permite al usuario final identificar en cuál clase recaen sus errores o la mayoría de éstos. Esta información no es proporcionada por los correctores usados en los procesadores de texto.

En la población de estudiantes analizada (ver Tabla 6.2) se identificó que los estudiantes de primer semestre cometen ligeramente más errores ortográficos y gramaticales que los estudiantes

de último semestre. En promedio, por cada 100 palabras, los estudiantes de primer semestre cometieron 0.08 errores contra 0.06 errores de estudiantes de los últimos semestres. Se observa que la diferencia del número de estudiantes entre los niveles bajo y alto en el dominio de los recursos gramaticales es la misma para los estudiantes de primero y últimos semestres: el 60 % de estudiantes recaen en un nivel bajo.

Tabla 6.2. Niveles de dominio de los recursos gramaticales de los estudiantes de licenciatura

Semestre	**Total de estudiantes**	**Nivel bajo**	**Nivel medio**	**Nivel alto**
Primer semestre	63	38	13	12
Último semestre	63	35	19	9
Totales	126	73	32	21

Se desarrolló una interfaz web con la que se espera que los usuarios mejoren su gramática. La herramienta puede ser accedida desde la dirección http://tecln.cenidet.edu.mx/grammarchecker/

A continuación, se listan algunas posibles aplicaciones del algoritmo:

1. Para instituciones educativas de nivel básico y medio superior, colegios, universidades: Se identificarán puntualmente los errores ortográficos y gramaticales los cuales podrían ser abordados en cursos o talleres de lectura y redacción para fomentar la correcta redacción de textos técnicos.

2. Para academias de enseñanza del idioma español. Para determinar el grado de dominio del idioma español en personas en las cuales no es su idioma materno.

6.1. Trabajos futuros

- **Detectar más reglas de uso para cada una de las categorías gramaticales:** Continuar con la detección de más reglas del uso correcto de todas las etiquetas *EAGLES* (adverbios, adjetivos, artículos, determinantes, nombres propios, pronombres, sustantivos, verbos, conjunciones y preposiciones) para integrarlo al sistema y hacerlo más preciso y amplié la cobertura en la detección de errores ortográficos y gramaticales.

- **Añadir más textos en español al corpus:**

  Continuar con la recopilación de textos en español que estén bien escritos, para incrementar el tamaño del corpus en español y reducir los errores estadísticos que ocurran en la etapa anterior.

- **Utilizar el código máquina generado por el sistema web para otros algoritmos y sistemas:**

  El código máquina que se genera en el sistema se utiliza en otros sistemas que requieran que se verifique la correcta utilización del lenguaje. Ejemplo: *Chatbots*, de traducción de texto, de extracción de información y de elaboración de resúmenes.

Referencias

Aguilar-Alconchel, M. (2004). Chomsky la gramática generativa. *Investigación y educación*, pp. 1–7.

Alberich, M. (2007). Procesamiento del lenguaje natural - guía introductoria. Recuperado el 19 de mayo de 2016, de Sopa de bits: http://www.sopadebits.com/wp-content/uploads/2011/03/4479-pln-1.0-20070630.pdf.

Alcina, A., Valero, E., y Rambla, E. (2009). *Terminología y Sociedad del conocimiento*. Peter Lang.

Araus-Gutíerrez, M., Chacón-Berruga, T., Cuesta-Martínez, P., Esgueva-Martínez, M., García-Macho, M., García-Page-Sánchez, M., Gómez-Manzano, P., y Martínez-Martín, F. (2010). *Curso básico de lengua española*. Centro de estudios Ramón Areces, S.A.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., y Padró, M. (2006). Freeling 1.3: Syntactic and semantic services in an open-source nlp library. *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC'06*, pp. 48–55.

Cirera, P. (2012). Introducción a las etiquetas eagles. Recuperado el 17 de abril de 2018, de LanguageTool en español: http://blade10.cs.upc.edu/freeling-old/doc/tagsets/tagset-es.html.

Domínguez, G. y Valcárcel, I. (2015). Diccionario de la rae en modo texto plano. Recuperado el 3 de junio de 2018, de Giusseppe Domínguez.net: https://www.giusseppe.net/blog/archivo/2015/10/29/diccionario-de-la-rae-en-modo-texto-plano/.

Echeverría-Arriagada, C. (2016). Attitudes and preferences of english-spanish translation students in relation to spanish grammatical features associated with frequency anglicisms. *Estudios filológicos*, 58:67–96.

Fernández-Fastuca, L. y Bressia, R. (2009). Definiciones y características de los principales tipos de texto. pp. 1–15.

Ferreira, A. y Kotz, G. (2010). Ele-tutor inteligente: Un analizador computacional para el tratamiento de errores gramaticales en español como lengua extranjera. *Revista signos*, 43(73):211–236.

Gelbukh, A. y Sidorov, G. (2006). *Procesamiento automático del español con enfoque en recursos léxicos grandes*. México: Instituto Politécnico Nacional.

Hernandez-Figueroa, Z., Carreras-Riudavets, F., y Rodriguez-Rodriguez, G. (2013). Automatic syllabification for spanish using lemmatization and derivation to solve the prefixs prominence issue. *Expert Systems with Applications*, 40:7122–7131.

Koehn, P. (2005). Europarl : A parallel corpus for statistical machine translation. *MT Summit*, 11:79–86.

Kotz, G. y Ferreira, A. (2012). La precisiǿn gramatical mediada por la tecnología: el análisis y tratamiento automático de errores. *Literatura y Lingüística*, 27(c):219–242.

LanguageTool.org (2016a). Languagetool supported languages. Recuperado el 21 de noviembre de 2016, de LanguageTool.org: https://www.languagetool.org/languages/.

LanguageTool.org (2016b). Reglas para languagetool. Recuperado el 30 de noviembre de 2016, de LanguageTool.org: http://community.languagetool.org/rule/list?lang=es.

Nordquist, R. (2018). Grammatical error definition and examples. Recuperado el 2 de junio de 2018, de ThoughtCo.com: https://www.thoughtco.com/grammatical-error-usage-1690911.

Oliva, M. y Serrano, M. (2010). Las bases cognitivas del estilo lingüístico. *Sociolinguistic Studies*.

Ramos, A. (2014). Las prácticas de evaluación docente y las habilidades de escritura requeridas en el nivel posgrado. *Innovación Educativa*, 14(66):147–176.

Real-Academia-Española (2010a). *Nueva gramática de la lengua española*. Espasa Calpe.

Real-Academia-Española (2010b). *Ortografía de la lengua española*. Espasa Calpe.

Rico, A. y Dimitrinka, N. (2015). Análisis de la competencia lingüístico-discursiva escrita de los alumnos de nuevo ingreso del grado de maestro en educación primaria. *Revista Signos*, 49(90):48–70.

San Mateo, A. (2016). Un corpus de bigramas utilizado como corrector ortográfico y gramatical destinado a hablantes nativos de español. *Revista signos*, 49:94–118.

Tausczik, Y. y Pennebaker, J. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

ThambiJose, F. (2014). Orthographic errors committed by sophomore students: A linguistic analysis. *Mediterranean Journal of Social Sciences*, 5:2439–2443.

Valera, J. (2013). Venelogia. Recuperado el 5 de mayo de 2016, de Venelogia: http://www.venelogia.com/archivos/7882/.

Vernon, S. y Alvarado, M. (2013). El desarrollo de la acentuación gráfica en niños y jóvenes mexicanos. *Revista mexicana de investigación educativa*, 18:141 – 157.

Apéndice A

Reglas gramaticales identificadas por *LanguageTool*

A continuación se mencionan las reglas gramaticales que identifica *LanguageTool* según LanguageTool.org (2016b).

- Concordancia
  - Ejemplos: 'las tres preceptos' etcétera
  - Ejemplos: 'los tres preguntas' etcétera
  - Ejemplo: 'alguna otro' -¿'alguna otra'
  - Errores de confusión. Ejemplo: *lo/los*
  - Concordancia singular en Determinante + nombre
  - Concordancia singular en «del» + nombre
  - Concordancia plural en Determinante + nombre
  - Concordancia singular Nombre + adjetivo
  - Concordancia plural Nombre + adjetivo
  - las casa
  - Concordancia femenino en Determinante + nombre
  - Palabras que empiezan por fonema A tónico
  - Concordancia masculino en Determinante + nombre
  - Concordancia masculino en «el» + nombre
  - Concordancia masculino en «del» + nombre
  - Concordancia femenino Nombre + adjetivo
  - Concordancia masculino Nombre + adjetivo
  - Concordancia «Cada una de los»
  - Concordancia «Cada una de las»
  - El agravante y el atenuante
  - Concordancia 1ª persona
  - Concordancia 2ª persona

- Concordancia 3ª persona
- Concordancia singular Nombre + verbo
- Concordancia plural Nombre + verbo

- Concordancia predictiva
  - Concordancia de sujeto y predicado en singular en oraciones atributivas
  - Concordancia de sujeto y predicado en número en oraciones atributivas
  - Concordancia de sujeto y predicado en femenino en oraciones atributivas
  - Concordancia de sujeto y predicado en masculino en oraciones atributivas
- Diversas
  - Repetición de una palabra
- Estilo
  - redundancia: orografía del terreno
  - Loor de multitudes
  - Redundancia neutro - femenino + masculino
  - Redundancia neutro - masculino + femenino
  - Redundancia neutro - femenino + masculino con partículas
  - Redundancia neutro - masculino + femenino con partículas
  - en relación a
  - relacionada/o/s a
  - relacionada/o/s al
  - *contracorriente, *a contra corriente
  - *contracorriente, *a contra corriente
  - *tal es así/tanto es así
- Gramática
  - Combinación imposible: preposición + verbo conjugado
  - *surgir/surtir efecto
  - en base a
  - «de gratis» (gratis)
  - Infinitivo tipo convencer + pron.pers + que
  - Estar + seguro + que
  - PP + verbo tipo alegrar + que
  - verbo tipo pensar + de que
  - Verbo tipo estar + AQ + de que
  - Verbo en 3a + de que
  - Verbo tipo insistir + de que

    - Adverbio de posición + posesivo (5)
    - no + imperativo
    - Uso incorrecto del plural del verbo haber
    - han *realizando/realizado
    - te se, me se
    - más + bueno/malo (mejor/peor)
    - A grosso modo –¿grosso modo
    - Sin en cambio –¿Sin embargo, en cambio
    - Si quiera–¿Siquiera
    - sino/si no (4)
- Mayúsculas y minúsculas
    - Comprobar si la frase se inicia con una letra mayúscula
- Ortografía (concepto)
    - Uso incorrecto de «haber» por «a ver»
    - e ante palabras empezando por i
    - cambio de o ante palabras empezando por o
    - afrentar dificultades o problemas (afrontar)
    - destornillarse de risa (desternillarse)
    - Contra más (cuanto más)
    - sin ecuánime (sine qua non)
- Ortografía (tipográficos)
    - uso/huso horario
    - a + participio
    - e+ participio
    - pronombre + e + participio
    - *ha/a + infinitivo
    - ah + infinitivo (a)
    - Hacer/ser, haz
    - ir ha + infinitivo (a)
    - ir + infinitivo (a)
    - Tu + verbo en 2a
    - Tu + conjunción
    - El + conjunción
    - Tu + partícula + verbo en 2a
    - el + verbo en 3a
    - Pronombre tú al final de la oración

    - Pronombre sin tilde al final de la oración
    - mi/mí
    - *esta/está + participio
    - lo se (sé)
    - se al final de la oración. (sé)
    - Verbo dé sin tilde al final de la oración
    - té sustantivo
    - Sustantivo té sin tilde al final de la oración
    - Se delante de ciertas palabas
    - Se + hacia es se hacía
    - aún/anun (7)
    - *apunto de/a punto de
- Posible error tipográfico
    - Posible error de ortografía
- Puntuación
    - Dos puntos o comas consecutivos
    - Paréntesis, comillas, signos de exclamación, interrogación y similares desparejados
- Tipografía
    - Espacios en blanco antes de coma y antes/después de paréntesis
    - Múltiples espacios en blanco
- Cambios de normas lingüísticas
    - sólo/solo
    - éste/este

Apéndice B

Palabras no identificadas por la librería *LanguageTool*

En la Tabla B.1 se muestra la lista de palabras no identificadas correctamente por la librería *LanguageTool*, y el total de errores en los documentos analizados en la fase 1 de las pruebas son:

Tabla B.1. Palabras no identificadas por la librería *LanguageTool*

Palabra incorrecta	**Palabra correcta**	**Total de errores**
Si	Sí	57
mas	más	37
esta	está	13
ademas	además	8
practica	práctica	8
practicas	prácticas	8
tenia	tenía	6
como	cómo	5
maquinas	máquinas	5
eh	he	4
mencione	mencioné	4
seria	sería	4
asi	así	3
asi mismo	así mismo	3
calculo	cálculo	3

Sigue en la página siguiente.

Palabra incorrecta	**Palabra correcta**	**Total de errores**
computo	cómputo	3
fabrica	fábrica	3
publicas	públicas	3
q’	que	3
ayudara	ayudará	2
cual’	cuál	2
mercadologicos	mercadológicos	2
que	qué	2
ala	a la	1
academicamente	académicamente	1
agrado	agradó	1
Agroecologia	Agroecología	1
agroinduztrial	agroindustrial	1
amplo	amplio	1
are	área	1
asta	hasta	1
busque	busqué	1
capas	capaz	1
capo	campo	1
de el	del	1
des	los	1
diferente	diferentes	1
el	en	1
empres	empresa	1
encargara	encargará	1
ensena	enseña	1
estada	estadía	1

Sigue en la página siguiente.

Palabra incorrecta	**Palabra correcta**	**Total de errores**
fabricas	fábricas	1
forjara	forjará	1
formula	fórmula	1
fuera	afuera	1
i	y	1
imagine	imaginé	1
llamo	llamó	1
maquina	máquina	1
marcara	marcará	1
mercadologicas	mercadológicas	1
metodo	método	1
necesitara	necesitaría	1
ó	o	1
optimisadores	optimizadores	1
optimo	óptimo	1
perdidas	pérdidas	1
porque	porqué	1
publica	pública	1
publico	público	1
q	que	1
sabia	sabía	1
sera	será	1
sobre peso	sobrepeso	1
solidas	sólidas	1
soluciona	solucionar	1
vé	ve	1
vera	verá	1

Sigue en la página siguiente.

Palabra incorrecta	**Palabra correcta**	**Total de errores**
ves	vez	1

Apéndice C

Cuestionario

Instrucciones: Redacta, por lo menos, una cuartilla en formato electrónico, de preferencia en block de notas, en la cual respondas a las siguientes preguntas. Tu escrito deberá ser claro y concreto.

1. ¿En qué consiste y cuál es el objetivo de la especialidad de la licenciatura que cursas actualmente?
2. ¿La carrera que actualmente estudias cumple con tus expectativas? Explica las razones de tu respuesta.
3. ¿Cuál es la utilidad de tu carrera en el campo laboral?
4. ¿Qué tipo de problemas resuelve un profesional de tu área? Por favor explica a detalle.