SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Validación automática del estilo de escritura en textos científicos de tipo tesis

presentada por

Ing. Claudia Lizeth Braulio Pérez

como requisito para la obtención del grado de

Maestría en Ciencias de la Computación

Director de tesis

Dr. Noé Alejandro Castro Sánchez

Codirector de tesis

Dr. Juan Gabriel González Serna

Cuernavaca, Morelos, México. Septiembre de 2018.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Morelos a 25 de junio del 2018
OFICIO No. DCC/208/2018

Asunto: Aceptación de documento de tesis

DR. GERARDO V. GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de la **Ing. Claudia Lizeth Braulio Pérez**, con número de control M16CE002, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado "**Validación automática del estilo de escritura en textos científicos de tipo Tesis**" 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

CO-DIRECTOR DE TESIS

Dr. Juan Gabriel González Serna
Doctor en Ciencias de la Computación
7820329

REVISOR 1

Dra. Andrea Magadán Salazar
Doctorado en Ciencias Computacionales
10654097

REVISOR 2

Dr. Máximo López Sánchez
Doctor en Ciencias de la Computación
7498547

C.p. M.T.I. María Elena Gómez Torres - 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

Cuernavaca, Mor., 20 de septiembre de 2018
OFICIO No. SAC/404/2018

Asunto: Autorización de impresión de tesis

ING. CLAUDIA LIZETH BRAULIO PÉREZ
CANDIDATA AL GRADO DE MAESTRA 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 **"Validación automática del estilo de escritura en textos científicos de tipo Tesis"**, 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"

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO SUBDIRECCIÓN ACADÉMICA

C.p. M.T.I. María Elena Gómez Torres.- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

Agradecimientos

Agradezco a Dios por permitirme llegar hasta esta etapa.

A mi madre por todo el sacrificio que ha hecho por mí, por apoyarme en cada momento y ser mi motivación en cada nuevo reto.

A mi padre, por ser un ejemplo a seguir y dejarme grandes enseñanzas para enfrentarme a la vida.

A mis hermanos, por apoyarme en cada una de las decisiones que he tomado y por estar conmigo a pesar de los obstáculos.

Al Dr. Noé Alejandro Castro Sánchez por confiar en mí, por sus consejos que siempre me ayudaron a mejorar profesional y académicamente.

Al Dr. Juan Gabriel González Serna, por sus consejos para ser mejor persona y estudiante, también por apoyarme en momentos difíciles.

A mis revisores la Dra. Andrea Magadán Salazar y al Dr. Máximo López Sánchez por sus observaciones y sugerencias que me ayudaron a mejorar mi trabajo de tesis.

A las excelentes personas que conocí en el transcurso de la maestría.

A mi amigo Samuel Rocha, por su amistad y compartir momentos complicados en el transcurso de la maestría.

Al personal de servicios escolares por su disposición de apoyarme cuando requerí de su ayuda para llevar a cabo el proceso de titulación.

A Liliana por apoyarme en todo momento con mis documentos, además de su apoyo con cada duda sobre el proceso de titulación.

A CONACYT por brindarme el apoyo y poder completar esta meta en mi vida.

A CENIDET por aceptarme en la maestría y ver en mi las aptitudes necesarias para cumplir con este grado de estudios.

Dedicatorias

A mi madre Alberta Pérez Allende, por ser un ejemplo de mujer y enseñarme a ser fuerte en la adversidad, por sus palabras de aliento para no rendirme y seguir adelante hasta conseguir el éxito, por su amor, cariño y comprensión en cada día de mi vida. También dedico esta Tesis a mis hermanos Andrés Braulio Pérez y Alberto Braulio Pérez por ser mi inspiración y el motor que me impulsaba a dar lo mejor de mí para ser un ejemplo digno para ellos, a mi padre Andrés Braulio Flores por darme las enseñanzas y valores para cumplir mis metas, sé que desde el cielo puede observar que hizo un buen trabajo y está orgulloso de mi.

Resumen

La escritura de textos científicos de tipo tesis es fundamental para la proyección de las ideas y una clara descripción del trabajo de investigación que se realiza. En su redacción se debe observar una correcta escritura que se rige por reglas gramaticales y de estilo, las cuales indicarán, entre otros elementos, el correcto uso de preposiciones y de la concordancia nominal y verbal, así como evitar jergas populares, vulgarismos, redundancia y pleonasmos. Desafortunadamente se ha detectado que los estudiantes tienen conocimientos deficientes en el área gramatical y en el uso adecuado del estilo que demanda la escritura a nivel postgrado.

El trabajo de esta tesis se orientó en la creación de una herramienta que valida la escritura de las tesis, atendiendo criterios gramaticales y de estilo. Dicha herramienta hace una evaluación de los siguientes capítulos: Introducción, Planteamiento del problema, Objetivos, Justificación, Alcances y limitaciones, Marco conceptual, Metodología de solución, Pruebas, Resultados y Conclusiones. Además de aplicar un análisis estilístico diferente a cada uno de ellos, pues siguen criterios que llegan a ser diferentes al resto.

Se identifican 7 tipos de errores gramaticales y 7 criterios estilísticos. La herramienta genera reportes que indican el número de errores encontrados, el tipo de error y su descripción. Su interfaz guía al usuario hacia la comprensión del error que se identifica al ofrecer una explicación de éste, así como en su corrección al proponer ejemplos y alternativas de solución para algunos de ellos.

Con los resultados de los reportes se observó que el error que se comete con más frecuencia es la redacción impersonal, la repetición de vocablos y las rimas.

Abstract

The writing of scientific texts of thesis type is fundamental for the projection of ideas and a clear description of the research work that is carried out. In its writing should be observed a correct writing that is governed by grammar rules and style, which will indicate, among other elements, the correct use of prepositions and nominal and verbal agreement, as well as avoid popular jargon, vulgarisms, redundancy and pleonasms. Unfortunately, it has been detected that postgraduate students have poor knowledge in the grammatical area and in the proper use of the style that writing demands at the graduate level.

The work of this thesis was oriented in the creation of a tool that validates the writing of the theses, attending grammatical and style criteria. This tool makes an evaluation of the following chapters:Introduction, Problem Statement, Justification, Scope and limitations, Conceptual framework, Solution methodology, Tests, Results y Conclusions. In addition to applying a stylistic analysis different to each of them, because each of them follows criteria that become different from the rest.

7 types of grammatical errors and 7 stylistic criteria are identified. The tool generates reports that indicate the number of errors found, the type of error and its description. Its interface guides the user towards understanding the error that is identified by offering an explanation of it, as well as correcting it by proposing examples and alternative solutions for some of them.

With the results of the reports it was observed that the most frequent error is the impersonal writing, the repetition of words and rhymes.

Contenido

Agradecimientos 3

Dedicatorias 4

Resumen 5

Abstract 6

Índice Figuras 10

Índice Tablas 12

Capítulo I 13

1.1 Introducción 14

1.2 Descripción del problema 15

1.3 Justificación 16

1.4 Objetivos 17

1.5 Alcances y Limitaciones 17

Capítulo II 19

2.1 Tesis 20

2.2 Gramática y tipos de fenómenos gramaticales 21

- **Gramática** 21
- **Preposición** 21
- **Dequeísmo** 22
- **Concordancia de género y número** 22

2.3 Fenómenos de Estilo 23

- **Jerga** 23
- **Riqueza léxica** 23
- **Rimas**23
- **Cacofonías** 23

2.4 Niveles de lenguaje 23

2.5 Lema 24

2.6 FreeLing 24

2.7 Part-of-Speech Tagging 25

Capítulo III 26

Estado del arte 27

3.1 Evaluación automatizada de la escritura científica [15] 27

3.2 Un identificador basado en reglas de gramática y estilo [16] 28
3.3 Reglas para la verificación gramática basadas en dependencias con *LanguageTool* [17] 30
3.4 Un informe sobre la evaluación automática de la escritura científica [18] 32
3.5 Sistema de corrección de errores gramaticales de POSTECH en CoNNL [19] 33
3.6 Combinación de herramientas de gramática y ortografía para la evaluación automática de escritura científica (AESW) Tarea compartida 2016 [20] 35
3.7 Introducción al Asistente de Escritura Científica (SWAN) Herramienta para evaluar la calidad de los manuscritos científicos [21] 36
3.8 Detección del nivel de dominio de recursos gramaticales en la redacción de textos técnicos de estudiantes de posgrado [22] 37
Capítulo IV 39
4.1 Reglas gramaticales 40
- **Reglas de preposiciones** 40
- **Reglas de preposiciones con verbos** 41
- **Reglas de Dequeísmo** 42
- **Reglas de Concordancia nominal y verbal** 44
- **Reglas del uso ante sustantivos femeninos que comienzan con /a/ o /ha/ tónica** 45

4.2 Reglas de estilo 47
- **Regla "Vicios del lenguaje (expresiones mal empleadas)"** 47
- **Regla "Uso de la persona gramatical"** 48
- **Reglas "Rimas"** 48
- **Regla "Cacofonías"** 48
- **Regla "Jergas populares"** 48
- **Regla Aberraciones lingüísticas** 49

4.3 Regla de similitud entre capítulos 49
Capítulo V 51
Metodología de Solución 52
5.1 FASE 1: Análisis de forma 53
5.2 FASE 2: Análisis de contenido 71
5.3 FASE 3: Generación de reportes 74
Capítulo VI 78
Aplicación Web 79
Capítulo VII 82
Pruebas 83

7.1 Pruebas Fase 1 83
- **Pruebas del módulo gramatical** 83
- **Pruebas del sub-módulo "Preposiciones"** 84
- **Pruebas del sub- módulo "Preposiciones con verbos"** 84
- **Pruebas del sub-módulo "Dequeísmo"** 85
- **Pruebas del sub-módulo "Concordancia nominal y verbal"** 86
- **Pruebas del sub-módulo "Anáfora"** 87
- **Pruebas del sub-módulo "Palabras tónicas"** 88
- **Pruebas del módulo de estilo** 89
- **Pruebas del sub-módulo "Vicios del lenguaje"** 89
- **Pruebas del sub-módulo "Uso de la persona gramatical"** 90
- **Pruebas del sub-módulo "Rimas"** 92
- **Pruebas del sub-módulo "Cacofonías"** 96
- **Pruebas del sub-módulo "Jergas populares"** 99
- **Pruebas del sub-módulo "Aberraciones lingüísticas"** 99

7.2 Pruebas Fase 2 100
- **Pruebas del sub-módulo "Repetición de vocablos"** 100
- **Pruebas del módulo "Identificación de similitud semántica entre capítulos"** 104

Capítulo VIII 114
Resultados 115
Capítulo IX 124
Conclusiones 125
Recursos generados 128
Anexo I 130
Referencias 146

Índice Figuras

Figura 1 Estructura de una tesis 20
Figura 2 Metodología general 52
Figura 3 Diagrama etapa 1 53
Figura 4 Diagrama etapa 2 54
Figura 5 Salida del sub-módulo "preposiciones" 54
Figura 6 Parámetros de entrada 55
Figura 7 Sugerencia de corrección 55
Figura 8 Salida del sub-módulo dequeísmo 56
Figura 9 Grafo en su fase inicial 58
Figura 10 Parámetros de entrada 59
Figura 11 Salida del sub-módulo de concordancia nominal y verbal 59
Figura 12 Salida del sub-módulo de concordancia nominal y verbal 60
Figura 13 Salida del sub-módulo "palabra tónica" 60
Figura 14 Módulo analizador de estilo 61
Figura 15 Metodología del sub-modulo "Vicios del lenguaje" 61
Figura 16 Salida del sub-módulo "Vicios del lenguaje" 62
Figura 17 Etiqueta morfosintáctica del verbo 62
Figura 18 Parámetros de entrada 63
Figura 19 Autómata del sub- módulo "Uso de la persona gramatical" 63
Figura 20 Salida del sub-módulo "Uso de la persona gramatical" 64
Figura 21 Autómata del sub- módulo "Uso de la persona gramatical" 64
Figura 22 Salida del sub-módulo "Uso de la persona gramatical" 65
Figura 23 Salida del sub-módulo "rimas" 66
Figura 24 Salida del sub-módulo "cacofonías" 68
Figura 25 Metodología del sub-módulo "rimas" y del sub-módulo "cacofonías" 68
Figura 26 Metodología del sub-módulo "jergas populares" 69
Figura 27 Salida del sub-módulo "jergas populares" 69
Figura 28 Salida del módulo "aberraciones lingüísticas" 70
Figura 29 Metodología del módulo "aberraciones lingüísticas" 70
Figura 30 Metodología del módulo "analizador de estilo" 71
Figura 31 Metodología del módulo "Riqueza léxica" 72
Figura 32 Salida del módulo "repetición de vocablos" 72
Figura 33 Metodología del módulo Analizador de riqueza léxica 73
Figura 34 Metodología del módulo "Identificación de relación entre capítulos" 73
Figura 35 Salida del módulo "Identificación de relación entre capítulos" 74
Figura 36 Metodología módulo Identificador entre capítulos 74
Figura 37 Reporte generado automáticamente del sub-módulo "preposiciones" 75
Figura 38 Reporte generado automáticamente del sub-módulo de dequeísmo 75
Figura 39 Reporte generado automáticamente del sub-módulo de concordancia nominal y verbal 75
Figura 40 Reporte sub-módulo "Palabras tónicas" 75
Figura 41 Reporte sub-módulo "vicios del lenguaje" 76
Figura 42 Reporte sub-módulo "Uso de la persona gramatical" 76
Figura 43 Reporte sub-módulo "Uso de la persona gramatical" 76
Figura 44 Reporte sub-módulo "rimas" 76

Figura 45 Reporte sub-módulo "cacofonías" 77
Figura 46 Reporte sub-módulo "jergas populares" 77
Figura 47 Reporte del módulo "aberraciones lingüísticas" 77
Figura 48 Vista principal de la aplicación Web 81
Figura 49 Tipos de errores analizados por la aplicación Web 81
Figura 50 Salida del sub-módulo "Preposiciones" 84
Figura 51 Salida del sub-módulo "Preposiciones con Verbos" 85
Figura 52 Salida del sub-módulo "Dequeísmo" 86
Figura 53 Salida del sub-módulo "Concordancia nominal y verbal" 87
Figura 54 Salida del sub-módulo "Anáforas" 88
Figura 55 Salida del sub-módulo "Tónica" 89
Figura 56 Salida del sub-módulo "Vicios del lenguaje" 90
Figura 57 Salida del sub-módulo "Uso de la persona gramatical" 91
Figura 58 Salida del sub-módulo "Uso de la persona gramatical" 91
Figura 59 Salida del sub-módulo "Uso de la persona gramatical" 92
Figura 60 Salida del sub-módulo "rimas" 95
Figura 61 Salida del sub-módulo "rimas" 98
Figura 62 Salida del sub-módulo "jergas populares" 99
Figura 63 Salida del módulo "aberraciones lingüísticas" 100
Figura 64 Análisis manual de repetición de vocablos 101
Figura 65 Salida del módulo "repetición de vocablos" 102
Figura 66 Determinación del parámetro "Mayoría" 107
Figura 67 Matriz de los porcentajes arrojados del algoritmo similitud coseno 113
Figura 68 Rango del intervalo de confianza 113
Figura 69 Errores de preposiciones analizados por la aplicación Web. 142
Figura 70 Errores de preposiciones analizados por la aplicación Web. 142
Figura 71 Errores de preposiciones con verbos analizados por la aplicación Web. 142
Figura 72 Errores de dequeísmo analizados por la aplicación Web. 142
Figura 73 Errores de concordancia nominal y verbal analizados por la aplicación Web. 143
Figura 74 Errores de anáforas analizados por la aplicación Web. 143
Figura 75 Errores de palabras tónicas analizados por la aplicación Web. 143
Figura 76 Errores de vicios del lenguaje analizados por la aplicación Web. 143
Figura 77 Errores de redacción impersonal analizados por la aplicación Web. 144
Figura 78 Errores de rimas analizados por la aplicación Web. 144
Figura 79 Errores de cacofonía analizados por la aplicación Web. 144
Figura 80 Errores de jergas populares analizados por la aplicación Web. 144
Figura 81 Errores de reiteración de vocablos analizados por la aplicación Web. 145
Figura 82 Errores de aberraciones lingüísticas analizados por la aplicación Web. 145

Índice Tablas

Tabla 1 Preposiciones mal empleadas 21
Tabla 2 Formas de uso de dequeísmo 22
Tabla 3 Tabla comparativa del estado del arte 38
Tabla 4 Lista de verbos y sus respectivas preposiciones 42
Tabla 5 Reglas de dequeísmo 42
Tabla 6 Verbos y dequeísmo 43
Tabla 7 Ejemplos del uso de dequeísmo 43
Tabla 8 Tipos de concordancia 44
Tabla 9 Reglas de concordancia nominal y verbal 44
Tabla 10 Ejemplos de sustantivos femeninos que comiencen por /a/ o /ha/ tónica 45
Tabla 11 Comparación de capítulos de una tesis 50
Tabla 12 Ejemplo de oración etiquetada 56
Tabla 13 Simplificación de etiquetas EAGLES 57
Tabla 14 Combinaciones de concordancia nominal y verbal 58
Tabla 15 Tipo de errores analizados por el sistema "Validador de textos científicos" 79
Tabla 16 Ejemplos de patrones de concordancia nominal y verbal 86
Tabla 17 Análisis de rimas 93
Tabla 18 Análisis de rimas en tesis 94
Tabla 19 Análisis de cacofonías 96
Tabla 20 Análisis de cacofonías en tesis 97
Tabla 21 Análisis manual de repetición de vocablos realizado por los hablantes nativos 103
Tabla 22 Comparación de capítulos de una tesis 104
Tabla 23 Comparación entre Titulo y Objetivos 106
Tabla 24 Comparación entre Planteamiento del problema y Objetivos 106
Tabla 25 Comparación entre Objetivos y Conclusión 107
Tabla 26 Datos para la evaluación de concordancia entre las comparaciones 108
Tabla 27 Acuerdo de evaluación por evaluador 109
Tabla 28 Acuerdo de evaluación total 110
Tabla 29 Datos para el cálculo de Kappa de Fleiss 111
Tabla 30 Interpretación de resultados de Kappa de Fleiss 112
Tabla 31 Interpretación de resultados de Kappa de Fleiss 118
Tabla 32 Distancia entre las palabras repetidas 119
Tabla 33 Comparación "Título – Objetivo" 121
Tabla 34 Comparación "Planteamiento del problema – Objetivo" 121
Tabla 35 Comparación "Objetivo - Conclusión" 122
Tabla 36 Precisión y cobertura de los diferentes fenómenos lingüísticos 123
Tabla 37 Rúbricas de evaluación 131
Tabla 38 Cálculo del Coeficiente de Alfa de Cronbach 132
Tabla 39 Rangos del coeficiente de Alfa de Cronbach 133
Tabla 40 Evaluaciones de capítulos 134
Tabla 41 Calculo de palabras correctas e incorrectas 135
Tabla 42 Parámetros de medición de calidad en la escritura 135
Tabla 43 Patrones de concordancia nominal y verbal 136

Capítulo I

Introducción

1.1 Introducción

La ciencia tiene un papel fundamental en el desarrollo de la humanidad, a lo largo del tiempo ha contribuido a que el hombre alcance el conocimiento del mundo que lo rodea. Gracias al avance científico es posible seguir desarrollando nuevas tecnologías,las cuales ayudan a mejorar la calidad de vida de los seres humanos. Los avances científicos son divulgados mediante textos científicos con el fin de propagar nuevos conocimientos en diversas disciplinas y poco a poco le han permitido al hombre dar respuesta a las inquietudes que han surgido.

Los textos científicos son aquellos cuyo contexto contienen de forma confiable todo el proceso que se requiere en una investigación con el objetivo de dar a conocer los resultados obtenidos [1], dependiendo de su estructura y del modo de escritura se clasifican en textos narrativos, descriptivos, expositivos, argumentativos e instructivso [2].

El texto científico se encuentra dentro de los textos expositivos, éstos generalmente son utilizados para demostrar los avances producidos durante una investigación. Así mismo, el texto científico se clasifica en artículos científicos, mural, tesina, monografías, informe científico, ensayo y tesis [3].

Como ya se mencionó, el principal objetivo de los textos científicos es transmitir nuevo conocimiento, por ello deben tener una redacción clara, sin ambigüedades y cubrir ciertos criterios gramaticales y de estilo.

Por esta razón, en este trabajo se expone el desarrollo de una herramienta que valida la escritura de los textos científicos de tipo tesis atendiendo criterios gramaticales y de estilo. Dicha validación es automática, lo cual permite detectar el mayor número de errores en el menor tiempo posible.

1.2 Descripción del problema

La Organización para la Cooperación y el Desarrollo Económicos (OCDE) estima que México se encuentra entre los 10 primeros países latinoamericanos con más graduados en postgrado, así mismo menciona que cerca del 22% de estudiantes mexicanos obtendrán un título de maestría en su vida [4], para ellos, los estudiantes tienen que escribir una tesis, en donde plasman el desarrollo y resultados de su investigación.

Un estudio realizado por la Universidad Autónoma Metropolitana, determina que de 99 alumnos solo el 1% de ellos cuenta con una buena redacción [5], esta situación es preocupante ya que el porcentaje de alumnos que tienen una buena escritura académica es muy bajo y por lo tanto la redacción de textos no cuenta con la calidad que se espera a nivel de postgrado.

Actualmente se han identificado varias problemáticas que presentan los estudiantes a la hora de redactar y plasmar sus ideas en una tesis de forma clara y coherente, entre las que se encuentran un deficiente léxico y el poco conocimiento de reglas gramaticales.

La problemática identificada consiste en que, durante el proceso de revisión de tesis del programa de Maestría en Ciencias de la Computación del Cenidet, se observan diversos errores cometidos en los escritos porque los estudiantes que las redactan no aplican correctamente las reglas gramaticales y de estilo, tales como el buen uso de preposiciones, dequeísmos, concordancia nominal y verbal, entre otras, por lo que se demerita la calidad de sus escritos. Tratándose de textos, la calidad conlleva la consideración de rasgos que sólo se pueden apreciar mediante la observación y juicios personales [6], por lo que basarse exclusivamente en mediciones cuantitativas no siempre resulta viable. Sin embargo, pueden considerarse algunas variables lingüísticas como indicadores de calidad, tales como la coherencia y la consistencia (la presencia de redundancias e incorrecto uso de concordancia nominal y verbal, etc.) (ver anexo I).

La solución de la problemática se aborda en este trabajo a través de la creación de una herramienta que valida la redacción de tesis considerando aspectos gramaticales y de estilo.

1.3 Justificación

Gracias a la escritura, el estudiante tiene la oportunidad de exponer el desarrollo y resultados de una investigación dada, con la finalidad de ser reconocido y aceptado por la comunidad científica.

Sin embargo, se han identificado varias problemáticas que presentan los estudiantes a la hora de redactar una tesis, esto se debe a que carecen de un léxico amplio y por ende su capacidad de utilizarlo para exponer sus ideas de forma clara y coherente es deficiente.

Un estudio realizado por la Universidad Autónoma Metropolitana, determina que de 99 alumnos solo el 1% de ellos cuenta con una buena redacción [5], esta situación es preocupante ya que el porcentaje de alumnos que cuenta con una buena escritura académica es baja y por lo tanto la redacción es deficiente y no permite la transmisión adecuada de la información de una tesis de postgrado.

Es por ello, que la importancia de realizar una validación automática de la redacción de tesis es necesaria para minimizar los errores en la redacción, y de esta forma asegurar que el contenido de ésta refleje de manera eficiente la investigación realizada.

1.4 Objetivos

• Objetivo general

El objetivo general de este trabajo consiste en desarrollar una herramienta que valide la redacción de textos científicos de tipo Tesis atendiendo criterios gramaticales y de estilo.

• Objetivos específicos

- Investigar el uso correcto del lenguaje en la redacción de trabajos de tesis.
- Desarrollar un módulo para el fragmentado de los capítulos de la tesis.
- Investigar las características de rimas en textos literarios y tesis.
- Investigar las características de cacofonías en textos literarios y tesis.
- Generar una lista de modismos, jergas populares y preposiciones.

1.5 Alcances y Limitaciones

• Alcances

- El sistema será capaz de detectar errores gramaticales y de estilo.
- El sistema creará sugerencia de sinónimos de aquellas palabras que presenten repeticiones.
- El análisis del texto considerará los requerimientos de redacción de cada una de las secciones de la tesis.
- El sistema realizará análisis a nivel morfológico y sintáctico.

- **Limitaciones**

- No se realizarán correcciones de coherencia.
- No se realizarán correcciones a nivel fonológico.
- El sistema funcionará solo en idioma español.

1.6 Estructura del documento

El contenido de la tesis se encuentra organizado en los siguientes capítulos:

- **Capítulo 2. Marco Conceptual:** En este capítulo se definen los principales conceptos para el tema de tesis desarrollado.
- **Capítulo 3. Estado del arte:** En este capítulo se presentan proyectos relacionados con la gramática y el estilo de los documentos científicos.
- **Capítulo 4. Reglas gramaticales y de estilo:** En este capítulo se presentan las reglas gramaticales y de estilo que se trabajaron en este trabajo de tesis.
- **Capítulo 5. Metodología de solución:** En este capítulo se describen las fases del método de solución.
- **Capítulo 6. Aplicación Web:** En este capítulo se muestra la aplicación web desarrollada en este trabajo de investigación.
- **Capítulo 7. Pruebas:** En este capítulo se presentan y explican detalladamente las pruebas aplicadas al sistema.
- **Capítulo 8. Resultados:** En este capítulo se exponen los resultados que arrojaron las pruebas aplicadas al sistema.
- **Capítulo 9. Conclusiones:** En este capítulo se exponen las conclusiones que se derivan de esta investigación.

Capítulo II

Marco conceptual

En el presente capítulo se presentan los conceptos utilizados como objeto de investigación de este trabajo, por lo tanto se incluyen términos tales como, tesis, gramática, sintaxis entre otros.

2.1 Tesis

La tesis es un texto científico en el cual se plasman los resultados de una investigación. El objetivo de una tesis es dar a conocer los resultados y conclusiones arrojados de una investigación exhaustiva y de alta complejidad [7].

La estructura de una tesis se muestra a continuación en la Figura 1.

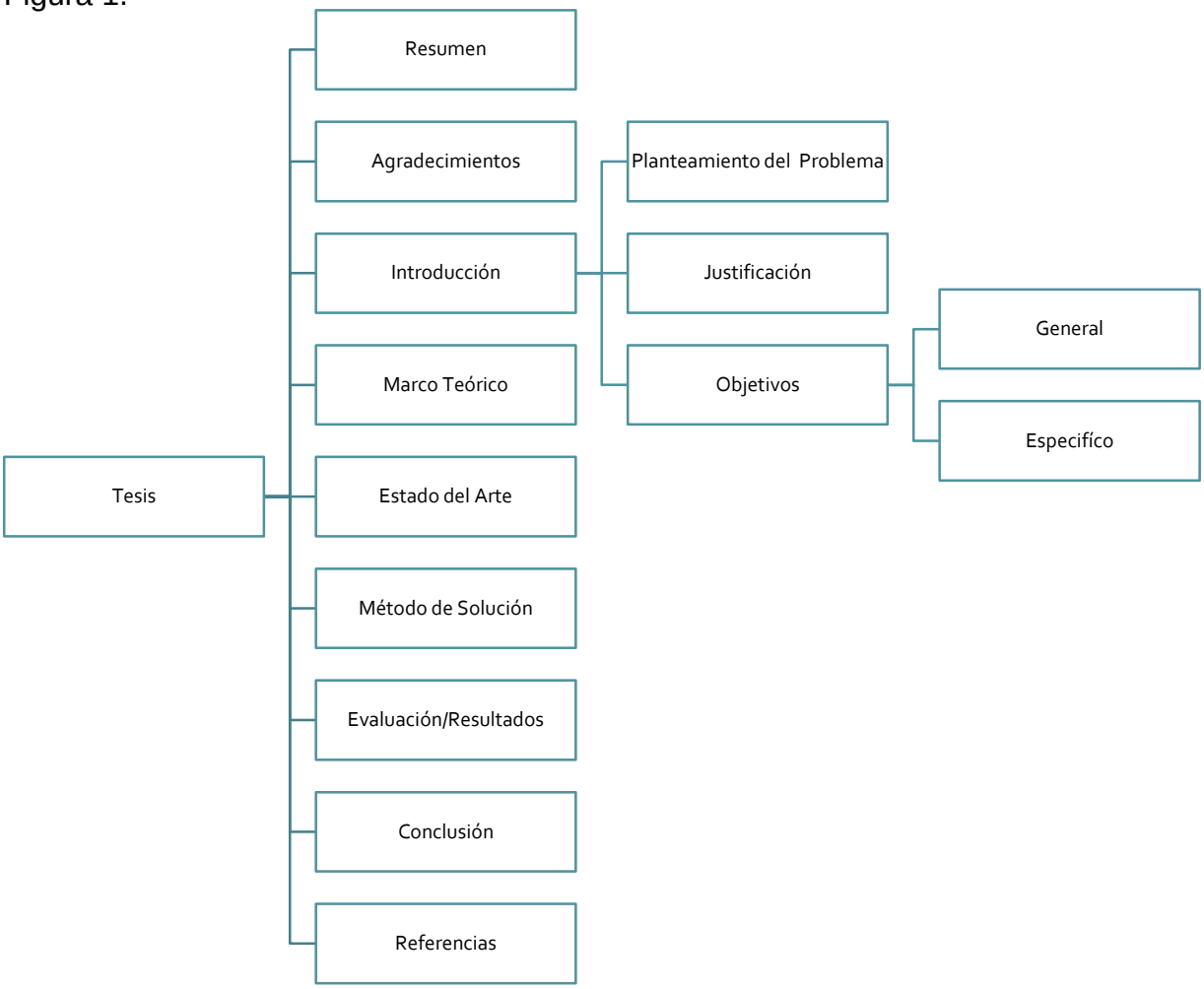


Figura 1 Estructura de una tesis

2.2 Gramática y tipos de fenómenos gramaticales

- **Gramática**

La gramática es una rama de la lingüística que estudia la estructura de las palabras, así como, la combinación de éstas para formar oraciones. Los elementos básicos de la gramática son el artículo, sustantivo, pronombre y verbo [8].

- **Preposición**

Sirve para unir términos, indicando una relación de dependencia. Señalan diversas relaciones como: dirección, lugar, tiempo, modo, limite, destino, cercanía, punto de partida, usanza, móvil, precio, entre otros [8]. Las preposiciones son: a, ante, hacia, hasta, cabe, con, por, pero, de, desde, sin, entre, en, tras, sobre, bajo, para, contra, según. Desgraciadamente existe una gran deficiencia en su uso a la hora de redactar textos.

En la Tabla 1 se muestra algunas preposiciones empleadas correcta e incorrectamente.

Tabla 1 Preposiciones mal empleadas

Forma correcta	Forma Incorrecta	Ejemplos
Cerca de	*Cerca a*	Mi casa queda *cerca de* la estación
Como consecuencia de	*En consecuencia a*	*Como consecuencia de* su actitud, ahora está en problemas.
Con relación a	*En relación a*	*Con relación a* ese asunto, Juan omitió opinar.
Junto con	*Junto a*	*Junto con* los otros veremos el partido.
Relacionado con	*Relacionado a*	Este hecho está *relacionado con* otros similares.
Con respecto a	*Respecto a*	Tenemos que hablar *con respecto a* tu situación.
a base de	*En base a*	*Sobre la base de* su nueva estrategia, el equipo saldrá victorioso.

- ## Dequeísmo

El dequeísmo consiste en utilizar de forma incorrecta la preposición “de” delante de la preposición “que”, por ejemplo:

“Dijo de que viene”.

Sin embargo, existen casos especiales en los verbos dudar, informar, advertir, avisar y cuidar. En este tipo de verbos está permitido usar las preposiciones “de” y “que” juntas, debido a que pueden utilizarse como complemento directo (sin preposiciones) o como un complemento de régimen (con preposiciones), por ejemplo, en la Tabla 2 se muestran los casos más comunes de dequeísmo

“*¿Dudas de que cumpla mi pacto?”*

Tabla 2 Formas de uso de dequeísmo

Ejemplo de dequeísmo	Forma incorrecta	Forma correcta
Cuando anteponemos la preposición “de” a una oración subordinada sustantiva de sujeto.	- Me pone contento de que estén cómodos	- Me pone contento que estén cómodos
Cuando ponemos la preposición “de” en locuciones conjuntivas que no la llevan.	- A no ser de que	- A no ser que
Cuando se usa la preposición “de” en vez de lo que el verbo precisa.	- Insistieron de que fuéramos	- Insistieron que fuéramos

- ## Concordancia de género y número

Según la Real Academia de la Lengua Española, concordancia es la coincidencia obligada de determinados accidentes gramaticales (género, número y persona) entre distintos elementos variables de la oración. Se pueden distinguir dos tipos de concordancia [7] :

- **Concordancia nominal** (coincidencia de género y número). Es la que establece el sustantivo con el artículo o los adjetivos que lo acompañan: “*La blanca paloma”*
- **Concordancia verbal** (coincidencia de número y persona). Es la que se establece entre el verbo y su sujeto “*Esos cantan muy bien”.*

2.3 Fenómenos de Estilo

- **Jerga**

La jerga se utiliza como una “lengua secreta” entre los individuos para comunicarse. La jerga está contenida por un léxico y una estructura lingüística especifica que son utilizados en diversas circunstancias [10].

- **Riqueza léxica**

Es la densidad de léxico en un texto, es decir, es la relación que existe entre el número de palabras totales y el número de palabras distintas que contiene el texto, por lo tanto, a mayor densidad léxica mayor es el número de palabras repetidas [11].

- **Rimas**

La rima se produce cuando existen la sucesión de palabras con igual terminación a poca distancia una de la otra. Por ejemplo:

"*En este trabajo se hace la identificación de un texto para su manipulación y extracción de resultados*"

- **Cacofonías**

La cacofonía según la RAE es la disonancia que resulta de la inarmónica combinación de los elementos acústicos de la palabra [8], se reconocen como cacofonías a la sucesión de palabras que empiezan con la misma sílaba o el mismo prefijo, por ejemplo: *“La **cla**sificación de los **cla**sificadores **clá**sicos”.*

2.4 Niveles de lenguaje

A continuación se describen los niveles de lenguaje que se abordan en este trabajo.

- **Morfología**

La morfología se encarga del estudio de las estructuras internas de las palabras [9] y como pueden modificarse. Su unidad mínima es el monema, existen monemas léxico o lexemas y los monemas gramaticales o morfema.

La morfología agrupa las palabras en clases de palabras denominadas categorías gramaticales de acuerdo a su estructura, función y significado.

- ### Sintaxis

La sintaxis estudia el orden y la relación que existen entre las palabras de una oración, es decir, las estructuras que se crean en función de la forma en que se ordenan o combinan las palabras [9]. Dentro de la sintaxis existen las oraciones enunciativas, oraciones exclamativas, oraciones desiderativas, oraciones interrogativas, oraciones dubitativas, oraciones imperativas o exhortativas.

- ### Semántica

Estudia el significado de las palabras y/o expresiones, así como, la relación de significado entre ellas [12]. Dentro de la semántica se encuentran los siguientes conceptos:

- **Homonimia:** Es la relación entre las mismas palabras, pero su origen y significado son diferentes.
- **Sinonimia:** Es la relación entre la palabra y el significado, se produce cuando dos o más palabras diferentes tiene el mismo significado.
- **Hiponimia/hiperonimia**: Es la relación entre significados

2.5 Lema

Se define como la representación de un conjunto de formas flexionadas derivadas de la misma palabra. Por lo tanto, el lema de las palabras: diré, dijo y dice; es decir.

Según la Real Academia Española (2016), el lema se refiere a la forma única común a todas las posibles variaciones de una misma palabra. En el caso de los sustantivos es utilizado el masculino singular de la palabra y en el caso de los verbos se utiliza la forma en infinitivo [13].

2.6 FreeLing

FreeLing es una librería desarrollada en el lenguaje de programación C++, esta herramienta permite realizar tareas de Procesamiento de Lenguaje Natural como: el etiquetado gramatical (etiquetado PoS por sus siglas en inglés, *Part of Speech*), reconocimiento y clasificación de entidades nombradas, desambiguación semántica, análisis morfológico, detección automática de idioma, etc.

Además, FreeLing es capaz de realizar todas las tareas ya mencionadas en los idiomas: español, inglés, portugués, italiano, francés, alemán, ruso, catalán, galés, croata, esloveno, etc [14].

2.7 Part-of-Speech Tagging

También conocido como PoS *tagging* o POST, se define como el proceso de asignación de etiquetas gramaticales a cada uno de los tokens que conforman un texto de acuerdo a su categoría gramatical [14]. Ejemplo:

El	sistema	ha	caducado
el	sistema	haber	caducar
DA0MS0	*NCMS000*	*VAIP3S0*	*VMP00SM*

Capítulo III

Estado del arte

Estado del arte

En esta sección se presentan las investigaciones relacionadas a este trabajo de tesis más relevantes en la literatura, en cada una de ellas se mencionan los objetivos y las herramientas que utilizaron para su desarrollo.

3.1 Evaluación automatizada de la escritura científica [15]

En este artículo se propone mejorar la calidad de la escritura científica mediante la correción gramatical de las oraciones, asi como la ortografía. Se basa en la mejora de la redacción de aquellos trabajos científicos que han sido elaborados por hablantes no nativos del inglés y por lo tanto no tienen la habilidad de escribir adecuadamente artículos en dicho idioma, esto genera que exista una gran demanda en la creación de herramientas para la escritura de textos científicos de calidad basados en sistemas computacionales. Así mismo, hace énfasis en el lenguaje académico y la lingüística académica científica.

El objetivo de este artículo es predecir si una determinada oración necesita cualquier tipo de edición para mejorar su redacción dentro del género de la escritura científica, puesto que detecta errores de gramática, reformulación, corrección ortográfica y correcciones de estructura de frase. Así como, analizar las características lingüísticas de la escritura científica con la finalidad de promover el desarrollo de herramientas automatizadas para la evaluaciónde escritura y con esto ayudar a los escritores en la redacción de artículos científicos.

Para cumplir con el objetivo establecido en este trabajo, primero se lleva a cabo una clasificación binaria, es decir, se examinan dos casos de decisiones, por un lado se hace la decisión binaria (falso o verdadera) y por otro la estimación probabilística (entre 0 y 1). Posteriormente, se extraerá el texto y se examinará de manera automatizada la evaluación de la escritura científica a nivel de oración, al final como salida se obtendrá un texto científico editado profesionalmente.

Cabe mencionar que para la extracción del texto, utilizaron la herramienta de código abierto tex2txt2, esta herramienta es utilizada para la conversión de LATEX a texto,

se desarrolló específicamente para esta tarea. La herramienta es independiente y no requiere de otras herramientas de procesamiento de látex o paquetes, su objetivo principal es extraer la información contextual correcta.

Algunos usos interesantes de las evaluaciones de calidad de nivel de la oración son los siguientes:

- Evaluación automatizada de la escritura en artículos científicos.
- Herramientas en la escritura de textos científicos en inglés
- Sentencias que necesitan mejorar la calidad de filtrado.

En resumen este artículo ayuda a la detección de las características de un buen lenguaje científico, así como, la mejora de los estándares aceptables para la redacción de artículos científicos.

3.2 Un identificador basado en reglas de gramática y estilo [16]

El objetivo de esta tesis es desarrollar un estilo de fuente abierta y gramatical para el idioma inglés. Los principales procesadores de texto de código abierto ofrecen corrección ortográfica, ninguno de ellos ofrece una función de corrector de estilo y de gramática, ni tampoco está disponible como un programa gratuito. Así, el resultado de esta tesis es un programa gratuito que puede ser utilizado como un complemento de un procesador de textos.

En esta tesis se describe un corrector de estilo y de gramática, en la cual se puede introducir un texto que después es analizado y como resultado devuelve una lista de posibles errores. Para detectar estos posibles errores a cada palabra del texto, se le asigna su etiqueta (*part-of-speech*) y cada frase se divide en fracciones, por ejemplo, en sintagmas nominales. A continuación, el texto se compara con todas las reglas de error que están ya predefinidas en el corrector, si una regla coincide se supone que en el texto existe un error en una determinada posición. Las reglas describen los errores como los patrones de palabras, cada regla incluye también una explicación del error, que se muestra al usuario.

El software planteado en esta tesis se basa en un sistema que se ha desarrollado anteriormente llamado “*Naber*”, por otra parte, para las tareas de etiquetado, reglas

gramaticales y de estilo se usó Python. En cuanto al sistema de reglas, éste se hará más potente, con la finalidad de utilizarlo para expresar reglas que describen los errores a nivel de frase y no sólo a nivel de palabra.

Por otro lado, la integración en los procesadores de texto será mejorado para que los errores pueden ser detectados en la marcha, es decir, durante la introducción de texto. Además de que ofrecerá una corrección que puede ser usado para reemplazar el texto correcto con un solo clic del ratón. Otra ventaja del software es que es lo suficientemente simple para que los usuarios puedan escribir sus propias reglas, sin embargo, es lo suficientemente potente como para obtener muchos errores típicos.

Entre las herramientas que utiliza, pueden enlistarse las siguientes:

- ***Ispell y Aspell*** ambos son correctores ortográficos de código abierto, la mayoría de los procesadores de texto hacen uso de estos programas. Por ejemplo, KWord proporciona una interfaz integrada para *Ispell y Aspell.*
- ***Estilo y dicción*** son dos comandos clásicos de Unix, los cuales calculan una serie de medidas de legibilidad como el Índice de *Flesch* o prueba de legibilidad de *Flesch-Kincaid*[1], también cuenta palabras, preguntas y frases largas (más de 30 palabras por defecto).
- ***EasyEnglish*** es un corrector gramatical desarrollado en IBM especialmente para los hablantes no nativos del idioma inglés.
- ***GramCheck*** es un corrector de estilo y gramática de español y griego, optimizado para los hablantes nativos de estas lenguas.
- ***FLAG*** (lenguaje flexible y revisión gramatical) es una plataforma para el desarrollo de aplicaciones de comprobación de lenguaje.

La mayoría de las normas se expresan en un simple formato XML, en el que se describen los errores y contiene un mensaje del mismo y ejemplos de frases. Aunque existen algunos errores que pueden ser demasiado complicados para ser expresado por las reglas en el archivo XML se pueden detectar mediante reglas escritas en Python, estas reglas también pueden ser fácilmente añadidos y no

[1] Prueba que busca conocer la facilidad de comprensión de un documento en inglés.

requieren ninguna modificación del código fuente existente. El software será alimentado con el Corpus Nacional Británicos (BNC), esto se realizará para probar errores reales, los cuales son extraídos de listas de correo y páginas web.

En conclusión, se han realizado varios proyectos científicos que trabajan en el estilo y la comprobación de la gramática, pero no hay ninguno disponible públicamente. Esta tesis y el software están disponible como software de código abierto en http://www.danielnaber.de/languagetool.

3.3 Reglas para la verificación gramática basadas en dependencias con *LanguageTool* [17]

El propósito de este artículo es diseñar una posible extensión para *LanguageTool*, es decir, se busca ampliar dicha fuente gramatical. Permite a los desarrolladores realizar dicha ampliación, ya que permite escribir las reglas gramaticales basadas en arboles de dependencia del analizador. Dichas reglas son importantes para realizar el análisis de enlaces de palabra-palabra, esto se hace son el fin de manejar una variedad de errores gramaticales incluyendo el uso indebido de artículos.

El proceso de la comprobación gramatical es un problema en el Procesamiento de Lenguaje Natural (PLN), se realizan con el propósito de encontrar errores gramaticales en un texto de entrada. Un corrector gramatical funciona normalmente en combinación con un corrector ortográfico, un módulo que detecta errores de ortografía en palabras individuales.

Los correctores ortográficos no pueden corregir defectos gramaticales, algunos paquetes que ofrecen un módulo gramatical son *Microsoft Office* o *WordPerfect Office*. Ciertos correctores gramaticales también están disponibles como paquetes de software adicionales o servicios en línea, ofrecido por empresas independientes. En cuanto, a las bibliotecas ortográficas abiertas, como JOrtho y GNU Aspell ya existen, y cualquier persona puede extender su propio software con sus capacidades.

El método que propone este trabajo tiene varias características atractivas. En primer lugar, las normas pueden ser fácilmente añadidas, modificadas o eliminadas; en segundo lugar, cada regla puede tener una amplia explicación correspondiente, útiles para el usuario final. Posteriormente, el sistema es fácilmente depurable, ya que sus decisiones se pueden remontar a una regla particular y, por último, las reglas pueden ser creadas por los lingüistas, que poseen pocos o nulos conocimientos de programación.

El algoritmo más simple gramática-estadístico consiste en analizar n-gramas de las cadenas de N palabras consecutivas [8]. Si una determinada cadena de palabra es común en el corpus de texto del patrón, se considera correcta. Además, los correctores gramaticales-estadísticos tienen sus propias ventajas y desventajas, pero su análisis está más allá del alcance de este artículo. También, utiliza la sentencia *Split* (*splitter* [divisor]) para determinar los límites de cada frase, permitiendo al usuario encontrar ciertos *tokens* al inicio o final de la oración.

La herramienta principal de este trabajo es *LanguageTool*, ésta usa dos bibliotecas para dividir y etiquetar texto de entrada, por ejemplo (*MXPOST* y *MXTERMINATOR*). *LanguageTool* detecta artículos “un y una” y detecta que deben ser empleados con un sustantivo en forma plural o plural. También utiliza dos analizadores *MaltParser* o *LDPar*, son de código abierto, *MaltParser* está escrito en Java, y de este modo se adapte mejor para el uso en combinación con la implementación actual de *LanguageTool*, también hecha con Java. Mientras tanto, la distribución *LDPar* contiene multiplataforma de código C ++, proporcionando una ejecución eficientemente compilable. Ambos programas de análisis se basan en el aprendizaje de la máquina, el analizador tiene primero que ser entrenados con una colección de oraciones correctamente analizados (un *Treebank*[2]). *MaltParser* y *LDPar* también comparten el mismo formato de los datos de entrada y de salida.

Para concluir, en este trabajo se diseñó e implementó un mecanismo de reglas del analizador de lenguaje natural, con ayuda de un módulo de comprobación gramatical basado en LanguageTool. Dicho mecanismo contiene una sintaxis que

[2] Es un corpus lingüístico en el que cada frase ha sido parseada, es decir, anotada con su estructura sintáctica.

permite la creación de reglas que analizan las dependencias de palabra-palabra en una frase dada.

3.4 Un informe sobre la evaluación automática de la escritura científica [18]

La evaluación automatizada de la escritura científica o AESW, es la tarea de identificar las oraciones que necesitan corrección para asegurar su pertinencia en una prosa científica. Este trabajo se centra en el problema de la identificación de frases en trabajos científicos que requieren edición. La principal motivación de este trabajo es promover el uso de herramientas de Procesamiento de Lenguaje Natural (PNL) para ayudar a escritores no nativos de inglés a mejorar la calidad de su escritura científica. Desde la perspectiva de la investigación, este esfuerzo tiene como objetivo promover un marco común y un conjunto de datos estándar para el desarrollo y pruebas del sistema de evaluación automática para el dominio de la escritura científica. El objetivo de esta investigación es estandarizar un conjunto de datos para el desarrollo de sistemas de evaluación automatizada para el domino de la escritura científica.

Utiliza la herramienta VTeX3 y el analizador *Stanford* (programa de análisis estadístico). El desarrollo del trabajo está basado en java, pero su reimplementación está en Python, cabe mencionar que está diseñado para el idioma inglés. Se llevan a cabo algoritmos en el que se analiza un corpus y mediante fórmulas se determina con una variable que oraciones pueden ser cándidas a posibles mejoras y con otra variable la que realmente si necesita ser modificada. Los algoritmos están basados en Decisión binaria y Estimación probabilística.

Para finalizar, se puntualiza que este trabajo se centra en el problema de identificar frases en obras científicas que requieran algún tipo de edición. La motivación del trabajo consiste en promover el uso de herramientas de PLN, con la finalidad de ayudar a escritores no nativos del idioma inglés a mejorar la calidad de la escritura científico en ellos. Cabe resaltar, que este trabajo ha servido de referencia para el estado de la técnica en la evaluación automática de la escritura científica, en la que

los resultados obtenidos demuestran que todavía hay margen de mejora. Teniendo en cuenta que la disponibilidad de que el conjunto de datos y las herramientas de evaluación faciliten el camino para futuros trabajos de investigación en esta área.

3.5 Sistema de corrección de errores gramaticales de POSTECH en CoNNL [19]

En este artículo se describe el sistema de corrección de errores gramaticales y ortográficos de la Universidad de Ciencia y Tecnología de Pohang, (POSTECH) en Corea del Sur. Para el desarrollo del sistema se proponen varios métodos para la corrección de errores, entre los cuales se encuentran, el método basado en reglas, el de vector de probabilidad de n-gramas y el basado en enrutadores.

En el artículo se menciona que la corrección automática de errores gramaticales, se emplea con frecuencia en estudiantes de inglés, por lo cual se han propuesto muchos métodos entre ellos, el basado en reglas (*Naber*), el de traducción automática estadística (*Brockett*), el de aprendizaje mecánico y n-gramas. Además de un método de meta-clasificación, con la finalidad de combinar un modelo de lenguaje y una clasificación especifica del error.

Los objetivos de este trabajo es realizar un algoritmo para corregir errores de preprocesamiento y ortografía, así como, corregir todos los errores gramaticales que se han tratado anteriormente, ya que en el 2012 el objetivo era corregir errores de artículos y de preposición, en el 2013 se corrigieron errores de artículo, preposición, número de sustantivo, la forma verbal y los errores de acuerdo sujeto-verbo.

El sistema está desarrollado en Python y utiliza la biblioteca Enchant para corregir los errores ortográficos, se utiliza un corpus de Google, el cual fue su principal recurso de corrección ya que cuenta con 1012 palabras de texto en ejecución. Por otro lado, para las pruebas de entrenamiento se extrajo información de un corpus escrito por estudiantes de la universidad de Singapur (NUCLE), cada texto candidato a ser analizado se evalúa con datos de desarrollo para extraer reglas de

alta precisión y marcos[3] de n-gramas. Cabe señalar que la utilización de un corpus bien formado ha funcionado de manera exitosa ya que están disponibles sin costo y poseen una gran cantidad de datos. También se utilizó *LanguageTool* para realizar correcciones de número de sustantivo y el analizador Stanford para extraer *part-of-speech* de la oración.

El método de corrección de errores gramaticales consta de varios casos, por ejemplo, como primer caso se hace el reencuentro de n-gramas, si el número de recuentro de n-gramas es bajo se asume que la palabra es errónea, en segundo caso, si la palabra de remplazo tiene más frecuencia que la palabra original, será una fuerte evidencia para realizar una corrección. En tercer lugar, dependiendo de la palabra candidata, la adaptación de marcos de n-gramas ayuda a corregir los errores de precisión, y por último, sólo se aplican métodos y reglas de alta precisión, si la precisión es menor de 30 % en la palabra no se realizará ninguna corrección.

En cuanto al proceso general, se corrigen los errores en el orden siguiente: como primer paso se realiza la tokenización para posteriormente hacer la corrección de errores ortográficos y errores de puntuación, para después ser analizado por medio n-gramas, y por último pasar por la corrección basada en enrutador para hacer la corrección gramatical.

En conclusión, en este artículo se describe el sistema de corrección de errores gramaticales y ortográficos POSTECH, se utilizó el corpus de google para detectar errores de ortografía y un método basado en reglas para errores gramaticales, así como, un enrutador para seleccionar la mejor secuencia de palabras y comparar la diferencia de puntaje de n-gramas entre el texto original y el texto candidato a analizar. Se pretende que en un fututo se pueda utilizar un método de aprendizaje para entrenar al enrutador de deferentes características.

[3] Secuencia de palabras alrededor de la posición objetivo.

3.6 Combinación de herramientas de gramática y ortografía para la evaluación automática de escritura científica (AESW) Tarea compartida 2016 [20]

En este artículo se presentan dos herramientas de código abierto para la detección de errores ortográficos y gramaticales*: After the Deadline* (corrección ortográfica, gramática y de estilo) y *LanguageTool* (corrección ortográfica y gramática).

El objetivo de este trabajo es promover el desarrollo de herramientas automatizadas de evaluación de escritura que ayuden a la tarea de escribir artículos científicos en el idioma inglés y predecir si una oración requiere edición para mejorar la redacción.

Este artículo describe que las herramientas para la corrección ortográfica y gramática existentes, se centran en textos académicos y debido a ellos generan pruebas para conocer cómo se comportan con este tipo de textos. En las pruebas realizadas utilizan las herramientas *After the Deadline* y *LanguageTool*, así como técnicas de unigramas, bigramas y trigramas.

En cuanto a la evaluación, se utilizó la herramienta Weka6, el conjunto de datos empleados fue seleccionado aleatoriamente de más de 9000 artículos de revistas de diferentes ámbitos entre ellos de informática.

Para el procesamiento, se dividieron los conjuntos de datos en archivos XML individuales, los cuales contienen 1000 oraciones cada uno, para realizar el aprendizaje automático de cada oración, se asignaron funciones de edición de verdadero y falso, si la oración tenía al menos una etiqueta se le asignaba la función de edición verdadero de lo contrario era falsa.

Los errores reportados se añaden al texto de entrada en forma de anotaciones, así como sus características, tipo de error y sugerencias de corrección. Como resultado final de este trabajo, demuestran que los experimentos realizados con ayuda de las herramientas pueden ayudar a evaluar la escritura académica pero no llegan a cubrir todas las correcciones de edición. Es por ello que, para trabajos futuros se planea clasificar falsos negativos y desarrollar más características para abarcar más tipos de errores. Sin embargo, a pesar de que se muestra que oración puede ser

problemática no se explica el por qué se marcó dicha oración o como podría mejorarse, lo cual es importante para los escritores académicos.

3.7 Introducción al Asistente de Escritura Científica (SWAN) Herramienta para evaluar la calidad de los manuscritos científicos [21]

Este trabajo es una tesis en la cual se presenta una herramienta para evaluar y mejorar la calidad de los escritos científicos en inglés, dicha herramienta lleva por nombre *Writing Assistant*). SWAN es una herramienta basada en el procesamiento de lenguaje natural, esta herramienta utiliza reglas, así como un conjunto de métricas, las cuales evalúan la calidad, la fluidez y la cohesión de un texto científico. La evaluación del texto se realiza por partes, es decir, se analizan por capítulos (título, resumen, introducción, conclusiones, etc).

El objetivo de esta tesis es proporcionar una visión detallada de las métricas necesarias para la creación de la herramienta SWAN. Cabe señalar que estas métricas fueron desarrolladas por capítulo.

Algunas de las métricas que trabaja SWAN consiste en determinan si existe consistencia entre los diferentes capítulos de la tesis, por ejemplo entre el *abstract* y el título, por ejemplo en el caso de la introducción lo que se busca es que tenga consistencia y no existan información innecesaria. Las pruebas de SWAN se llevaron a cabo con usuarios, los cuales evaluaron su experiencia al interactuar con la interfaz de la herramienta, dando como resultados que, para los usuarios, SWAN es una herramienta útil que mejora la calidad de la escritura científica.

En conclusión, SWAM ayuda a que un artículo científico contenga una estructura informativa, lógica, consistente, limpia, clara, concisa, además ayuda a que el título y *abstract* estén relacionados.

3.8 Detección del nivel de dominio de recursos gramaticales en la redacción de textos técnicos de estudiantes de posgrado [22]

En el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), se cuenta con una tesis en desarrollo, la cual lleva por nombre “Detección del nivel de dominio de recursos gramaticales en la redacción de textos técnicos de estudiantes de posgrado” realizada por el estudiante Leonel González Vidales, esta tesis está muy relacionada a este trabajo de investigación ya que en ella se realiza un estudio acerca de los errores ortográficos y gramaticales que más comenten los estudiantes de posgrado.

Su objetivo primordial es diseñar un algoritmo que permita identificar el nivel de dominio de los recursos gramaticales en la redacción de textos técnicos de aspirantes al posgrado. En dicha tesis se están utilizando herramientas como *LanguageTool* para realizar correcciones ortográficas y gramaticales, esta herramienta es de código abierto y está disponible para varios idiomas, entre ellos, inglés, español francés, entre otros. También se han realizado búsquedas de recursos léxicos con la finalidad de desarrollar reglas que permitan la creación de un módulo de análisis gramatical, así como la búsqueda de listas de la Real Academia Española de la Lengua.

Hasta el momento la tesis realizada por Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), cuenta con un prototipo funcional para la revisión ortográfica y gramatical, dicho módulo está desarrollado en el lenguaje de programación Java. Para realizar pruebas se utilizaron propuesta de tesis y reportes de avances, estos documentos fueron sometidos a un análisis ortográfico y gramatical, los resultados de este análisis son visualizados en un reporte en el cual se muestran los errores detectados.

En la Tabla 3 se realiza una comparación de los trabajos correspondientes al estado del arte y el trabajo desarrollado durante esta investigación.

Tabla 3 Tabla comparativa del estado del arte

Tema	Idioma	Corrección gramatical	Corrección de estilo	Ámbito de interacción	Herramientas utilizadas
Evaluación automática de la escritura científica	Inglés	✓	✗	Artículos científicos	▪ Tex2txt ▪ Decisión binaria ▪ Probabilidad estadística
Un identificador de gramática y estilo basado en reglas	Inglés	✓	✓	Artículos científicos	▪ Ispell y Aspell ▪ EasyEnglish ▪ GramCheck ▪ LanguageTool ▪ Style y Diction
Reglas para la verificación gramática basada en dependencias con LanguageTool	Inglés	✓	✗	Artículos científicos	▪ JOrtho ▪ Aspell ▪ LanguageTool
Informe sobre la evaluación automática de la escritura científica	Inglés	✓	✗	Artículos científicos	▪ Analizador Stanford ▪ Decisión binaria ▪ Decisión probabilística
Sistema de corrección de errores gramaticales de POSTECH en CoNNL	Inglés	✓	✗	No específica	▪ Analizador Stanford ▪ LanguageTool ▪ Vector de probabilidad de n-gramas ▪ Enrutadores
Combinación de herramientas de gramática y ortografía para la evaluación automática de escritura científica (AESW) Tarea compartida 2016	Inglés	✓	✓	Artículos científicos	▪ After the Deadline ▪ LanguageTool
Introducción al Asistente de Escritura Científica (SWAN) Herramienta para evaluar la calidad de la Manuscritos científicos	Inglés	✗	✓	Artículos científicos	▪ Stanford Parser ▪ Stanford Part-Of-Speech (POS)
Validación automática de textos científicos estilo tesis	Español	✓	✓	Tesis	▪ FreeLing ▪ Aspell ▪ LanguageTool

Como se aprecia en la tabla, el trabajo de tesis desarrollado que lleva por nombre "Validación automática de textos científicos estilo tesis, abarca los fenómenos lingüísticos gramaticales y de estilo, además de que está disponible en el idioma español, lo cual hasta el momento no se encuentra en ningún trabajo de investigación.

Capítulo IV

Reglas gramaticales y de estilo

A continuación se describen las reglas gramaticales y de estilo que sigue el idioma español, las cuales fueron las bases para el desarrollo de la herramienta que se desarrolló en este trabajo.

4.1 Reglas gramaticales

La gramática es una disciplina que se encarga de explicar la forma en que los elementos de la lengua se enlazan para formar textos y también se encarga de analizar los significados de estas combinaciones [23].

La gramática se divide en tres partes; la primera se encarga de la fonología, la segunda corresponde a la morfología y la tercera está relacionada con la sintaxis, ésta última se encarga de la estructura de los enunciados [23].

Existen diversos criterios gramaticales que deben tomarse en cuenta a la hora de redactar una tesis. En este trabajo se tomaron en cuenta criterios gramaticales de concordancia nominal y verbal, así como de preposiciones y dequeísmo, los cuales se presentan a continuación.

- **Reglas de preposiciones**

La preposición es una palabra invariable y átona que tiene la función de introducir un sustantivo o un grupo nominal, con éste se forma un complemento que depende sintácticamente de otro elemento del enunciado. Las preposiciones del español actual son las siguientes: a, ante, bajo, con, contra, de, desde, durante, en, entre, hacia, hasta, mediante, para, por, según, sin, sobre, tras [24].

Las preposiciones son unidades dependientes que incrementan a los sustantivos, a los adjetivos o a los adverbios como índices explícitos de las funciones de tal manera que las palabras cumplen correctamente en la oración [18]. Por ejemplo, la preposición “de”, en el siguiente enunciado: *María habla de la escuela*

Enlaza el núcleo verbal habla con su término adyacente “*la escuela”* de forma que de *“la escuela”* queda como objeto preposicional del verbo.

O bien, en el siguiente enunciado: *Aún persiste el recuerdo de la escuela*

El segmento “*de la escuela”* es adyacente del sustantivo precedente “*el recuerdo”.*

Como se aprecia en los ejemplos anteriores, la preposición por sí sola no cumple función alguna dentro del texto, solo sirve como índice del segmento en el que encuentra integrada. Aunque cabe mencionar que algunas preposiciones están dotadas de un significado más o menos explícito pero que va dependiendo del contexto en el que se encuentran, es decir, las preposiciones pueden comportase como un valor léxico.

En otras palabras, las preposiciones son palabras que ayudan a relacionar los diferentes elementos de una oración, es por ello, que su buen uso es vital para transmitir de forma correcta un mensaje.

Según Orlando Gabriel Cáceres Ramírez, experto en ortografía y redacción, el mal uso de las preposiciones se encuentra entre los errores más comunes en la gramática española [24]. Estudios realizados por la Universidad Autónoma de Barcelona, señalan que cuando en una oración se encuentra el conector "que" y un artículo, la preposición debe anteponerse al artículo y no al conector [25].

En el Anexo 1 se muestran los errores más frecuentes del uso de preposiciones, así como el uso correcto de conectores con preposiciones.

- ## Reglas de preposiciones con verbos

Existen determinadas preposiciones que deben ir acompañadas de un verbo en específico, es decir, que la preposición debe obedecer la noción léxica del núcleo verbal que se encuentra adyacente a ella [24]. Así por ejemplo los verbos, hablar, pensar y contar requieren respectivamente las siguientes preposiciones *de*, *en* y *con*.

El significado de la raíz verbal y de la preposición se vuelve una función obligatoria, como se muestra en los siguientes enunciados:

- *Piensa <u>en</u> su familia*
- *Contamos <u>con</u> el dinero*

Es por lo anterior que se realizó la indagación de los verbos que deben ir acompañados de una preposición específica. En la Tabla 4 se muestran algunos verbos y sus respectivas preposiciones.

Tabla 4 Lista de verbos y sus respectivas preposiciones

Verbo	Preposición
acertar acercar acudir	a
contar divertir casar	con
olvidar parar pasar	de
consentir consistir convenir	en

- ## Reglas de Dequeísmo

En la gramática del castellano se encuentra un error llamado dequeísmo, el cual consiste en el uso indebido de la preposición “de” delante de la conjunción “que” cuando la preposición no viene exigida por ninguna palabra en el enunciado [26].

A continuación en la Tabla 5 se presentan las reglas recabadas que se implementaron:

Tabla 5 Reglas de dequeísmo

Reglas de Dequeísmo	
1	Cuando se antepone la preposición *de* a una oración subordinada sustantiva de complemento directo. Esto ocurre, sobre todo, con verbos de «pensamiento» (*pensar, opinar, creer, considerar,* etc.), de «habla» (*decir, comunicar, exponer,* etc.), de «temor» (*temer,* etc.) y de «percepción» (*ver, oír,* etc.). • *Pienso QUE conseguiremos ganar el campeonato* • *Me dijeron QUE se iban a cambiar de casa*
2	Cuando se antepone la preposición de a una oración subordinada que ejerce funciones de atributo en oraciones copulativas con el verbo ser: • *Mi intención es QUE participemos todos*
3	Cuando se inserta la preposición de en locuciones conjuntivas que no la llevan: • *a no ser QUE*
4	Cuando se usa la preposición de en lugar de la que realmente exige el verbo: • *Insistieron en QUE fuéramos con ellos*
5	Los verbos “advertir, avisar, cuidar, dudar e informar”, en sus acepciones más comunes, pueden construirse de dos formas: • *Me informó DE QUE venía* • *Me informó QUE venía*

Sin embargo, existen verbos en los que se permite el dequeísmo, en la Tabla 6 se muestran estos verbos.

Tabla 6 Verbos y dequeísmo

Verbo	Dequeísmo
Pensar	✗
Opinar	✗
Creer	✗
Considerar	✗
Decir	✗
Comunicar	✗
Exponer	✗
Temer	✗
Ver	✗
Oír	✗
Ser	✗
Insistir	✗
Fijar	✗
Advertir	✓
Avisar	✓
Cuidar	✓
Dudar	✓
Informar	✓

En Tabla 7 se muestra algunos ejemplos del uso correcto de esta regla gramatical:

Tabla 7 Ejemplos del uso de dequeísmo

Verbo	Ejemplo	Estado
Decir	Me dijo ***de que*** venía	Incorrecto
Avisar	Me aviso ***de que*** venía	Correcto
Avisar	Me aviso ***que*** venía	Correcto
Sentir	Siento ***de que*** ganaremos	Incorrecto

- ## Reglas de Concordancia nominal y verbal

La concordancia es la coincidencia obligatoria entre los elementos variables (género, número y persona) de una oración. Existen dos tipos de concordancia, los cuales se muestran en la Tabla 8, en donde las letras “M” y “F” hacen referencia al género “Masculino” y “Femenino” respectivamente y las letras “S” y “P” hacen referencia al número “Singular” y “Plural”.

Tabla 8 Tipos de concordancia

Concordancia nominal						
Artículo		**Sustantivo**		**Adjetivo**		
El	**(M)**	niño	**(M)**	bonito	**(M)**	
La	**(F)**	niña	**(F)**	bonita	**(F)**	
Concordancia verbal						
Artículo		**Sustantivo**		**Adjetivo**		**Verbo**	
El (M)		niño	**(M)**	bonito	**(M)**	canta	**(S)**
Las	**(F)**	niñas	**(F)**	bonitas	**(F)**	cantan	**(P)**

Se investigaron las reglas que deben respetarse para lograr la concordancia de género (masculino, femenino) y número (singular, plural) y la concordancia verbal.

En la Tabla 9 se presentan las reglas generales de concordancia que se implementaron:

Tabla 9 Reglas de concordancia nominal y verbal

Reglas de Concordancia nominal y verbal	
1	La coordinación del determinante con el sustantivo. • ***La aplicación***
2	La coordinación del determinante o pronombre con el adjetivo. • ***La*** *aplicación* ***compleja*** • ***Ésta*** *es* ***compleja***
3	La coordinación del sustantivo con el verbo. • *La* ***aplicación es*** *compleja*
4	La coordinación del determinante con el verbo. • ***La*** *aplicación* ***es*** *compleja*
5	La coordinación del sustantivo con el adjetivo. • *La* ***aplicación*** *es* ***compleja***

- **Reglas del uso ante sustantivos femeninos que comienzan con /a/ o /ha/ tónica**

Cuando se presenta un sustantivo femenino que comience por */a/* o */ha/*, el artículo femenino "*la*" que lo acompaña se convierte obligatoriamente en el artículo masculino "*el*" , por ejemplo: la palabra "águila" es un sustantivo femenino y se pensaría que debe ir acompañado por el artículo femenino "la", pero la "á" con la que empieza la palabra es tónica por lo tanto el artículo debe cambiarse a "el" [26]. Quedando de la siguiente manera:

La águila El águila

Esta regla solo funciona cuando se trata de términos en singular, es decir, en el caso del plural quedaría de la siguiente manera:

El águila Las águilas

Sin embargo, cuando el artículo acompaña a topónimos femeninos que empiecen por la letra /a/ tónica, el uso de esta regla es variante. Por ejemplo con los nombres de los continentes se emplea la forma /el/ :

La África El África

La Asia El Asia

En la Tabla 10 se muestra algunos ejemplos del uso del artículo con palabras que empiezan por la letra /a/ y /ha/ tónica.

Tabla 10 Ejemplos de sustantivos femeninos que comiencen por /a/ o /ha/ tónica

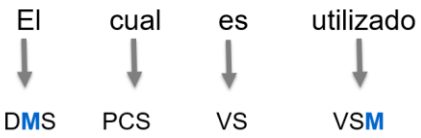
✗	✓
La águila	El águila
La aula	El aula
La hacha	El hacha
La hambre	El hambre

Existen excepciones en las que la frecuencia de uso de los hablantes, ha inhibido esta regla, por ejemplo: La palabra árbitra (derivado de árbitro) ha sido una palabra modificada por los hablantes, los cuales dicen “la árbitra” y no “el árbitra”. Es por ello, que esta norma puede rechazarse dependiendo de las tradiciones de los hablantes.

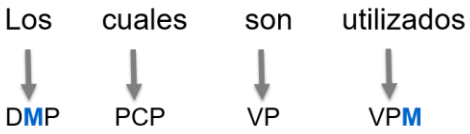
Otras reglas de concordancia son las siguientes:

1.- Pronombres relativos

Cuando se tiene un pronombre relativo de género común, acompañada de un determinante y verbo se tiene que evaluar el género del determinante y del verbo. En el siguiente ejemplo se tiene un pronombre común y el determinante que lo precede “*el*” es “masculino”/” singular” por lo tanto, el verbo que lo acompaña debe estar en “masculino”/” singular”

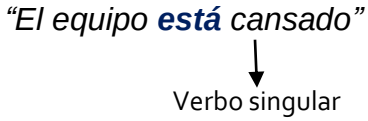


Mientras tanto en el siguiente ejemplo el pronombre común está en plural y el determinante que lo precede está en “masculino” /” singular”, por lo tanto, el verbo que lo también debe estar en “masculino” /” singular”.



2.- Verbo de sustantivos colectivos

Se creó una regla para reconocer que un verbo colectivo como “comunidad”, “grupo, “equipo”, debe ir acompañado de un verbo en singular. Por ejemplo:



3.- Sustantivo acompañado de la preposición “*de”*

También se creó una regla, que evalúa el sustantivo que se encuentra antes de la preposición “de”, para determinar el género y numero del sustantivo y con esto determinar el verbo que debe acompañarlo. Por ejemplo:

*“El grupo de estudiantes **es** pequeño”*

Verbo singular

4.2 Reglas de estilo

La redacción de una tesis debe cumplir con parámetros de estilo específicos, ya que son importantes para su correcta transmisión [29]. Por lo tanto, la tesis debe estar escrita de manera clara y concisa, los parámetros de estilo que se tomaron en cuenta para desarrollar el módulo “analizador de estilo” se muestran en los siguientes capítulos.

- **Regla “Vicios del lenguaje (expresiones mal empleadas)”**

Existen vicios del lenguaje que se encargan de pronunciar o escribir mal algunas palabras, con la creencia de que tienen cierto significado cuando en realidad tienen otro. A este tipo de palabras que sufren estas modificaciones se les conoce como *barbarismos* [24], algunos ejemplos son “antier”, “pobrísimo”, “nuevísimo”, entre otros.

Los barbarismos son tan usados que a veces terminan siendo aceptados, lo mismo sucede con las palabras o expresiones lingüísticas que un determinado idioma toma de otra lengua extranjera que se emplean innecesariamente de manera frecuente por las personas, este tipo de palabras son conocidas como *extranjerismos* [27].

Debido a la frecuencia de su uso, la RAE en el *Lista Panhispánico de Dudas* [26] ya se han aceptado algunos extranjerismos dados a su equivalente en español como: software, sport, pizza, ojalá, corsé, jazz, blues, rugby, entre otros, pero existen otros los cuales tienen usos innecesarios tales como: “*email”, “show”, “look”, “link”, “online”, “parking”, “ok”, “play”, “clik”, “short”, “happy”, “bay”, “tablet”, etc*. Es por ello, que el experto en lingüística Orlando Cáceres Ramírez, aseguró que la RAE

“desaconseja el uso de este tipo de extranjerismo ya que existen palabras que tienen el mismo significado en español”.

Asimismo, también existen expresiones o palabras que son incorrectas porque posen alteraciones morfológicas, sintácticas o fonéticas, dichas palabras son rechazadas por hablantes cultos y por lo tanto no son admitidos en la redacción de textos científicos a este tipo de palabras se les conoce como *vulgarismo* [28] y por último se encuentran los errores de repetición de palabras/vocablos que son innecesarias, este tipo de errores se conoce como *pleonasmos*. En el anexo 1 se muestran las listas de cada uno de los tipos de vicios del lenguaje que se han mencionado.

- ## Regla “Uso de la persona gramatical”

La escritura de una Tesis debe estar en modo impersonal, por lo tanto, se recomienda el uso de la tercera persona [9]. Mientras tanto el capítulo de objetivos debe contener verbos en infinitivo.

- ## Reglas “Rimas”

La rima se produce cuando existen la sucesión de palabras con igual terminación a poca distancia una de la otra. Por ejemplo: *“El diput**ado** no ha disput**ado** lo que se había pronunci**ado**”.*

- ## Regla “Cacofonías”

La cacofonía según la RAE es la disonancia que resulta de la inarmónica combinación de los elementos acústicos de la palabra [8], se reconocen como cacofonías a la sucesión de palabras que empiezan con la misma sílaba o el mismo prefijo. Por ejemplo: *“La **cla**sificación de los **cla**sificadores **clá**sicos”.*

- ## Regla “Jergas populares”

Las jergas populares son expresiones que se originan por la edad o los diferentes grupos y sirven para diferenciar un grupo de otro [10]. Algunos ejemplos son: “claridoso”, “compa”, “cuentear”, “chance”, “mero”, “checar”, “destrampe”, entre otros.

- **Regla Aberraciones lingüísticas**

El contenido de una tesis debe ser clara, precisa y concisa para transmitir de forma eficiente los resultados de dicha investigación es por ello que se llevó a cabo un estudio acerca de las aberraciones lingüísticas más frecuentes en la redacción de una Tesis, este tipo de deficiencias afecta la claridad del texto científico [11]. Dichas aberraciones se tomaron de algunos portales web especializados en lingüística y de artículos enfocados a las deficiencias del lenguaje en español. Se tomaron en cuenta dos tipos de aberraciones lingüísticas, las cuales se describen a continuación:

- Verbosidad: Este tipo de vicio hace que la escritura se vuelva tediosa ya que no es concisa ni precisa. Algunos ejemplos son: “A pesar del hecho que” el cual puede sustituirse únicamente por la palabra “aunque”, otro ejemplo es “es capaz de” por “puede”.
- Vocabulario rebuscado: en este tipo de error del lenguaje genera que la comunicación no sea precisa, ya que son palabras poco comunes y es mejor usar el sinónimo de esas palabras en la escritura del texto, por ejemplo: “hipodigmo = muestra”, “espurio= falso”, “proclive = propenso”, entre otros.

4.3 Regla de similitud entre capítulos

En la elaboración de tesis existen dos requisitos para lograr su comprensión y elegancia. El primero de ellos se refiere al uso apropiado del lenguaje y a la organización del texto, mientras que el segundo trata del fondo, el cual abarca la armonía de todas las ideas, la demostración del análisis que debe conducir a las conclusiones, la profundidad que trata de la esencia del problema y la originalidad, la cual se logra mediante el análisis de los intentos realizados por otros investigadores o por el propio investigador de resolver problemas. Así bien, con la finalidad de obtener una mejor calidad en la tesis se recomiendan las relaciones

entre los siguientes capítulos[4]: “Título” - “Objetivo”, “Objetivo” - “Planteamiento del problema” y “Objetivo” - “Conclusión”.

En la Tabla 11 se muestran estas relaciones, así como la explicación de su importancia.

Tabla 11 Comparación de capítulos de una tesis

Relación	Importancia
“Título – Objetivo”	El título representa la esencia de la investigación de forma concreta y generalmente se define brevemente el objetivo de ésta.
“Objetivo” - “Planteamiento del problema”	El planteamiento del problema debe contener los argumentos que describan el interés de la investigación a realizar, contiene información acerca los planteamientos específicos que identifican lo que se desea lograr con el proyecto de investigación. También se describe la finalidad de la investigación y en ésta se hace referencia al objetivo general y objetivos específicos los cuales corresponden al enfoque propuesto. Es decir, los objetivos responden con exactitud a la definición del problema.
“Objetivo-Conclusión”	La conclusión es la argumentación fundamentada de la problemática y en ésta se dan a conocer los resultados obtenidos. También, se da a conocer si se cumplieron los objetivos planteados.

[4] http://www.upv.es/laboluz/master/seminario/textos/umberto_eco.pdf

Capítulo V

Metodología de solución

Metodología de Solución

Para el desarrollo de este trabajo de investigación se implementó una metodología de solución compuesta por tres fases: Fase 1: Análisis de forma, Fase 2: Análisis de contenido y Fase 3: Generación de reportes. A continuación, la Figura 2 se ejemplifica en un diagrama la metodología utilizada.

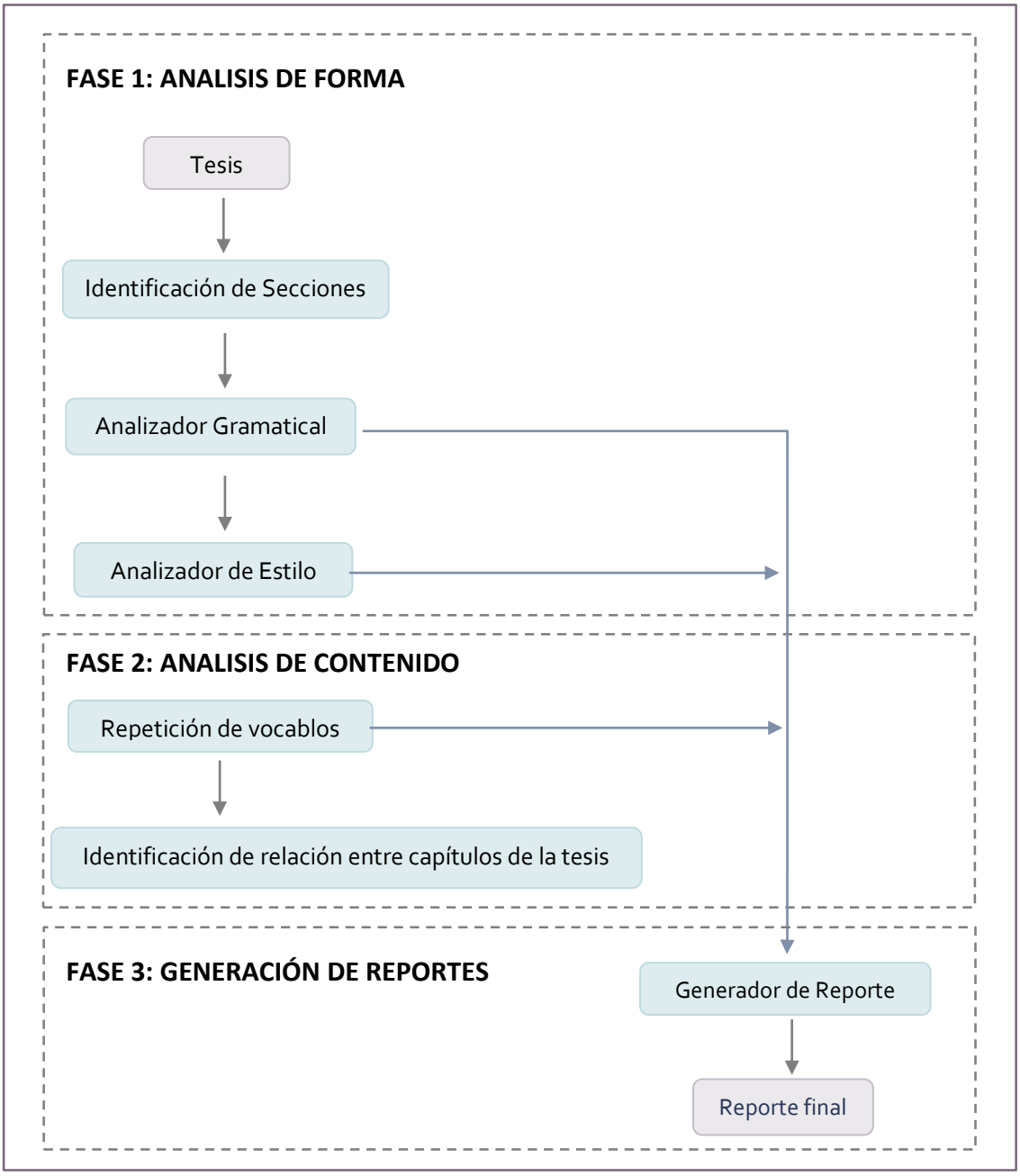


Figura 2 Metodología general

5.1 FASE 1: Análisis de forma

La primera fase de la metodología está definida por tres etapas, cada una de ellas cumple un propósito determinado con la finalidad de identificar el mayor número de errores gramaticales y de estilo, así como el reporte de éstos.

- ### Etapa 1: Identificación de Secciones

Una tesis, como ya se mencionó en los capítulos anteriores, consiste de diversos capítulos y cada uno de ellos tiene diferentes métricas que cubrir, es por ello, que en esta etapa se hace la segmentación de capítulos para aplicar dichas métricas.

Una vez hecha la segmentación, se realizó el etiquetado de oraciones para poder manipular los datos con ayuda de la herramienta *Freeling*[5] como resultado final se obtiene la segmentación de capítulos en un archivo de texto plano con las palabras etiquetadas, en la Figura 3 se aprecia el diagrama de esta etapa.

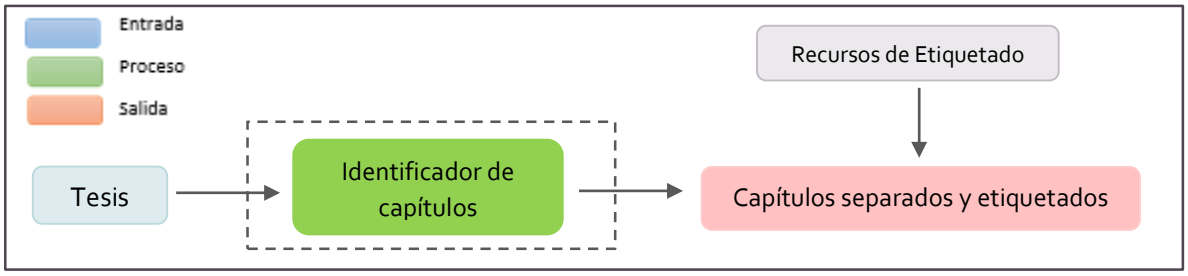


Figura 3 Diagrama etapa 1

- ### Etapa 2: Analizador gramatical

Esta etapa, consiste en el desarrollo de un analizador gramatical por capítulos de la tesis, el cual recibe como entrada los capítulos segmentados y detecta los siguientes fenómenos gramaticales.

- Preposiciones
- Preposiciones con verbos
- Dequeísmo
- Discordancia nominal y verbal
- Anáforas
- Palabras tónicas

[5] http://nlp.lsi.upc.edu/freeling/node/1

En la Figura 4 se observa el diagrama general de la metodología empleada.

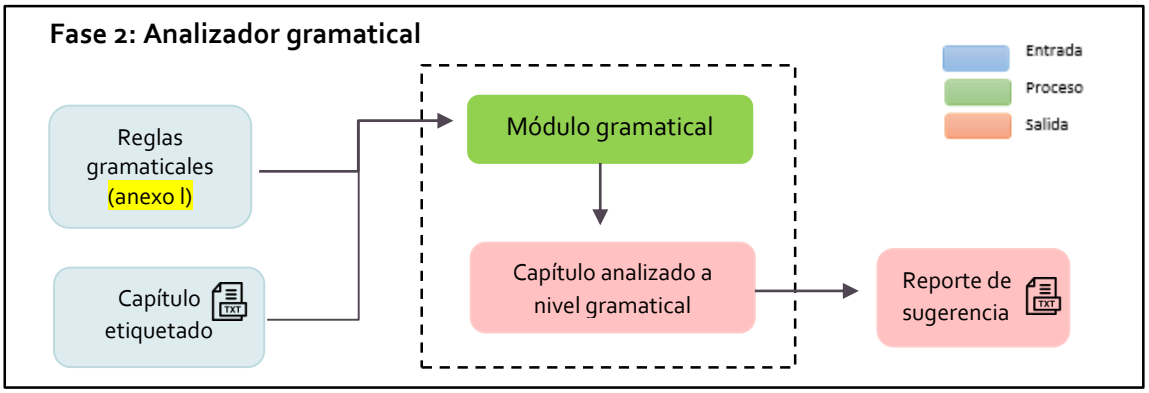


Figura 4 Diagrama etapa 2

A continuación se describen cada uno de los sub-módulos dedicados a analizar cada uno de los criterios gramaticales antes mencionados.

- ## Sub-Módulo de Preposiciones

Como primer paso se desarrolló el sub-módulo dedicado a la detección de errores en el uso de preposiciones, para esto se cuenta con dos listas de preposiciones, una incorrecta (lista 1) y otra correcta (lista 2). Los datos de entrada a este sub-módulo son en texto plano (txt), para su posterior análisis, el cual consiste en identificar si existe en el texto alguna similitud de la lista 1 y sustituirla por la forma correcta de la lista 2.

Finalmente como salida se muestra el texto remplazado correctamente, como se aprecia en la Figura 5, en color rojo es la forma incorrecta del uso de preposiciones y en color azul el remplazo correcto.

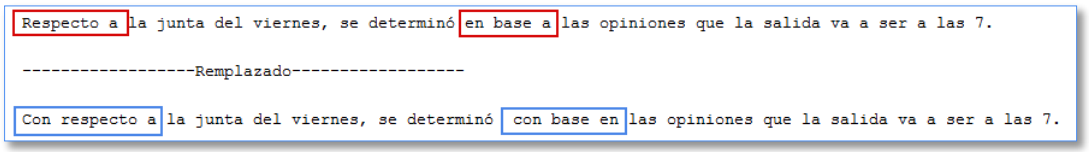


Figura 5 Salida del sub-módulo "preposiciones"

Como se mencionó en la sección anterior, existe una regla que tiene que ver con las preposiciones y el verbo que las acompaña, para dicha regla se desarrolló un

sub- módulo que se encarga de analizar el lema del verbo y su preposición, para ello se recolectaron y clasificaron los verbos dependiendo de las preposiciones que deben emplearse.

Una vez clasificados los verbos se crearon 4 listas, la elaboración de las listas se hizo con la finalidad de realizar comparaciones entre los verbos de la lista y los verbos del texto a analizar. La entrada a este módulo es un archivo en formato txt, el cual contiene tres parámetros (palabra, lema y etiqueta). En este caso se toma el segundo parámetro, es decir, el lema de las palabras (Figura 6).

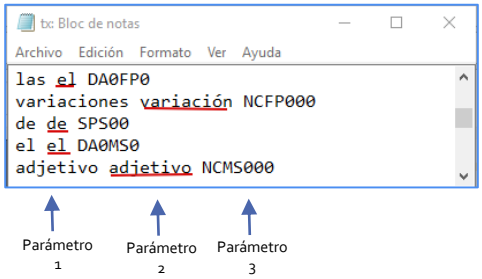


Figura 6 Parámetros de entrada

Este sub-módulo analiza los verbos del archivo (txt) y las preposiciones que estén enseguida de ellos, si dicho verbo emplea una preposición errónea se imprime una alerta y la sugerencia de corrección, como se aprecia la Figura 7.

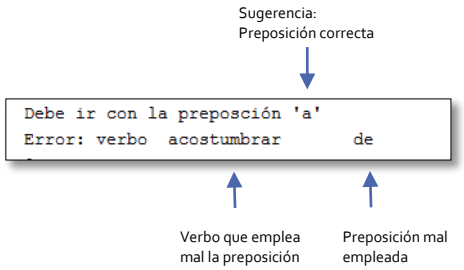


Figura 7 Sugerencia de corrección

- ## Sub-módulo Dequeísmo

Para esta regla se desarrolló un módulo que consiste en detectar si en el texto se encuentra la preposición “de” y el conector “que” juntos.

Para este módulo se cuenta con una lista de verbos que no permiten el dequeísmo. La entrada a este módulo es un archivo en formato txt, el cual contiene tres

parámetros (palabra, lema y etiqueta), en este caso al igual que en el sub-módulo de preposiciones se toma el segundo parámetro.

Posteriormente se analiza el texto de entrada (txt), para identificar si se encuentra “de que”, de ser cierto se analiza el verbo que se antepone y se compara con la lista de verbos, si el verbo analizado se encuentra en la lista se emite un mensaje con la sugerencia de corrección, como se observa en la Figura 8.

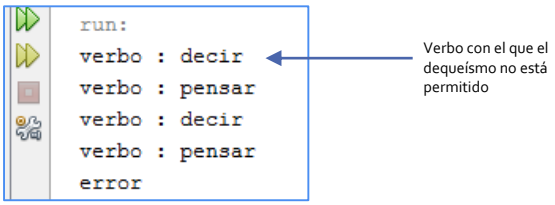


Figura 8 Salida del sub-módulo dequeísmo

- ## Sub-módulo Concordancia de género y número

Para este sub-modulo se desarrolló un autómata, el cual está alimentado de reglas y patrones de concordancia nominal y verbal (Anexo 1). A continuación se explica la forma en que se construyó el autómata para llevar a cabo este módulo.

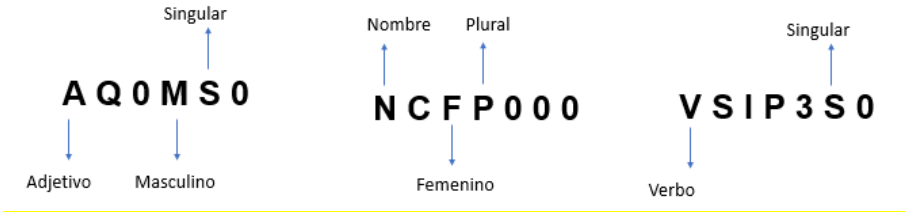
Primeramente, se mostrará un ejemplo de concordancia nominal y verbal. Esta regla consiste en la presencia de dos o más sustantivos en una oración, usando las etiquetas EAGLES ver Tabla 12.

Tabla 12 Ejemplo de oración etiquetada

La	aplicación	,	el	sistema	y	el	código	son	sencillos
DA0FS0	NCFS000	FC	DA0MS0	NCMS000	cc	DA0MS0	NCMS0	VSIP3P0	AQ0MP0

Como se observa en el ejemplo anterior, aunque se tenga femenino y masculino en singular, el verbo y adjetivo que los describen están en masculino y plural. Por lo anterior, como primer paso se realizó la clasificación de las etiquetas, dividiéndolas en grupos y simplificándolas dejando sólo la información morfosintáctica necesaria, para esto se realizaron patrones con la ayuda de expresiones regulares.

La simplificación de las etiquetas se realizó con la finalidad de manipular fácilmente las etiquetas y de obtener sola la información en cuanto al género y número, como se puede ver a continuación en las siguientes etiquetas:



A continuación en la Tabla 13 se muestran algunas etiquetas que se utilizaron en el autómata.

Tabla 13 Simplificación de etiquetas EAGLES

Simplificación de etiquetas EAGLES				
Tipo	Género	Número	Etiqueta	Simplificación
Determinantes	Masculino	Singular	DD0MS0 DP1MSP DI0MS0 DA0MS0	DDMS DPMS DIMS DAMS
		Plural	DD0MP0 DP1MPP DI0MP0 DA0MP0	DDMP DPMP DIMP DAMP
	Femenino	Singular	DD0FS0 DP1FSP DI0FS0 DA0FS0	DDFS DPFS DIFS DAFS
		Plural	DD0FP0 DP1FPP DI0FP0 DI0FP0 DA0FP0	DDFP DPFP DIFP DAFP
	Común	Singular	DI0CS0	DICS
		Plural	DI0CP0	DICP
Nombres	Masculino	Singular	NCMS000	NMS
		Plural	NCFS000	NFS
		Invariable	NCMN00	NMN
	Femenino	Singular	NCFS000	NFS
		Plural	NCFP000	NFP
		Invariable	NCFN000	NFN
	Común	Singular	NCCS000	NCS
		Plural	NCCP000	NCP
	Propio		NP000G0 NP000O0 NP000P0	NP
Verbos		Singular	VSII1S0 VSIP3S0 VMIP1S0	VS
		Plural	VSIP3P0 VSIC3P0 VMIP3P0	VP

Como siguiente paso se determinaron las posibles combinaciones que pueden surgir en la concordancia, en la Tabla 14 se muestran algunas de estas combinaciones [13].

Tabla 14 Combinaciones de concordancia nominal y verbal

DAMS	+	NCMS							
DAMS	+	NCMS	+	AQMS					
DAFS	+	NCFS							
DAFS	+	NCFS	+	AQFS					
DAMP	+	NCMP							
DAMP	+	NCMP	+	AQMP					
DAFP	+	NCFP							
DAFP	+	NCFP	+	AQFP					
DAMS	+	NCMS	+	DAFS	+	NCFS	CC	VP	AQMP
DAFS	+	NCFS	+	DAMS	+	NCMS	CC	VP	AQMP

Como se aprecia en la tabla anterior, todas las combinaciones que se tomaron en cuenta comienzan con etiquetas de tipo determinante, es por eso que el autómata cuenta con 4 entradas, que son determinante en masculino en (singular y plural) y femenino en (singular y plural). Una vez detectando el tipo de determinante, el autómata analiza la siguiente palabra y recorre los nodos dependiendo de la información morfosintáctica de la etiqueta. En la Figura 9 se muestra un el primer prototipo de grafo para analizar esta regla.

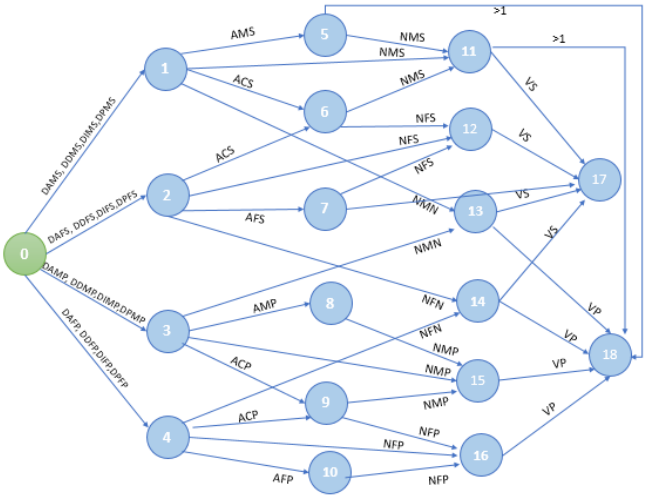


Figura 9 Grafo en su fase inicial

La entrada a este módulo es un archivo en formato txt, el cual contiene tres parámetros (palabra, lema y etiqueta), en este caso se tomó el parámetro 3 (Figura 10).

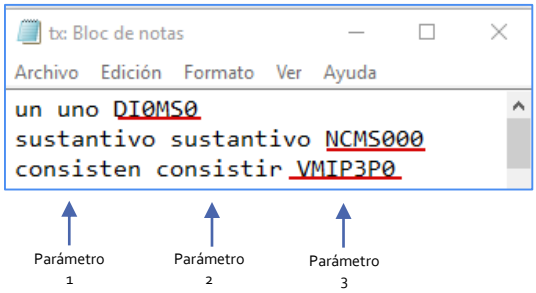


Figura 10 Parámetros de entrada

Posteriormente se analizan las oraciones del archivo de entrada (etiquetadas). Existen cuatro entradas al autómata (anexo 1), las cuales son: Determinante Masculino Singular (DMS), Determinante Femenino Singular (DFS), Determinante Masculino Plural (DMP) y Determinante Femenino Plural (DFS).

Posteriormente las etiquetas que son determinantes entran al primer nodo del autómata y empieza el recorrido hasta generar la salida dependiendo del patrón en que se clasifique la oración analizada, si la oración es correcta se imprime un mensaje de "correcto", en cambio si se presenta una inconsistencia de concordancia nominal o verbal se manda una advertencia de incorrecto.

La salida de este módulo concordancia nominal y verbal se aprecia en la Figura 11 (oración correcta) y Figura 12 (oración incorrecta).

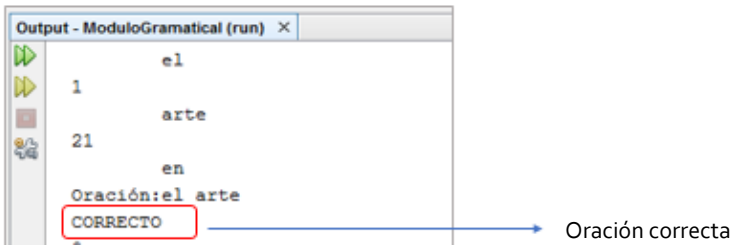


Figura 11 Salida del sub-módulo de concordancia nominal y verbal

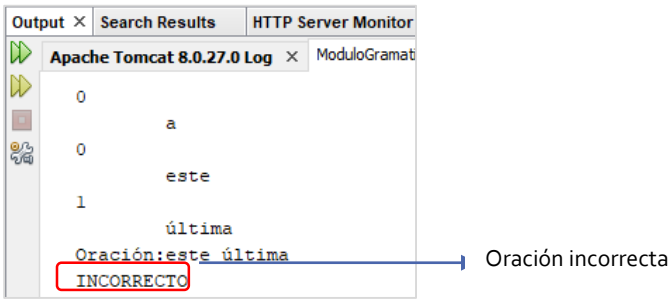


Figura 12 Salida del sub-módulo de concordancia nominal y verbal

También se desarrolló un sub-módulo en el que se identifican las palabras que empiecen por /a/ o /ha/ tónica y detecte que se esté utilizando el artículo adecuado. Para ello, se creó una lista de palabras que poseen la peculiaridad de que la primera letra empieza por /a/ o /ha/ tónica.

La entrada a este módulo es un archivo en formato plano (txt) etiquetado, del cual se extraen las palabras y estas son comparadas con las palabras que se encuentran en la lista, al detectar una igualdad entre las palabras de la lista y las del texto de entrada, se analiza el artículo que la antepone y se emite una advertencia de error si se llega a encontrar una inconsistencia. En la Figura 13 se muestran los datos de salida de este sub-módulo.

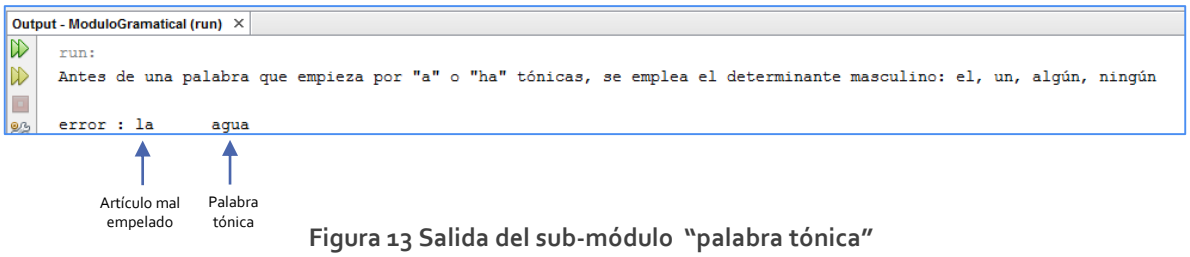


Figura 13 Salida del sub-módulo "palabra tónica"

- ## Etapa 3: Analizador de estilo

La etapa 3, consiste en la creación de un módulo llamado “Analizador de estilo”, el cual está dividido en seis sub-módulos y está desarrollado en el lenguaje de programación Java (Figura 14).

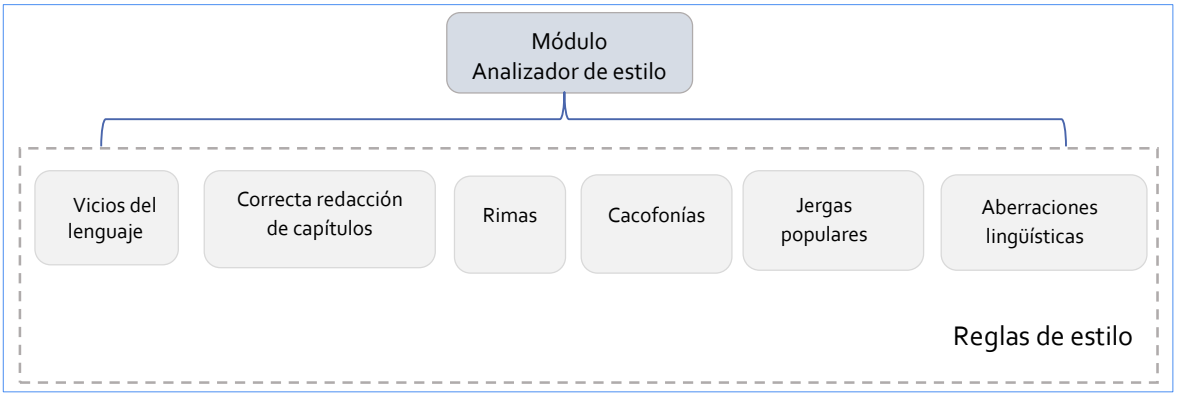


Figura 14 Módulo analizador de estilo

A continuación se explica de forma detallada cada uno de los sub-módulo que componen el Analizador de estilo.

- ## Sub-módulo “Vicios del lenguaje (expresiones mal empleadas)”

Este sub-módulo se encarga de detectar y analizar aquellas expresiones como vulgarismo, extranjerismos, barbarismo y pleonasmos. Cuenta con dos listas, una incorrecta (lista 1) y otra correcta (lista 2) por cada uno de los tipos de expresiones mal empleadas. El análisis de este sub-módulo consiste en detectar alguna similitud de la lista 1 y sustituirla por la forma correcta de la lista 2, en la Figura 15 se muestran la metodología para realizar la sustitución correcta.

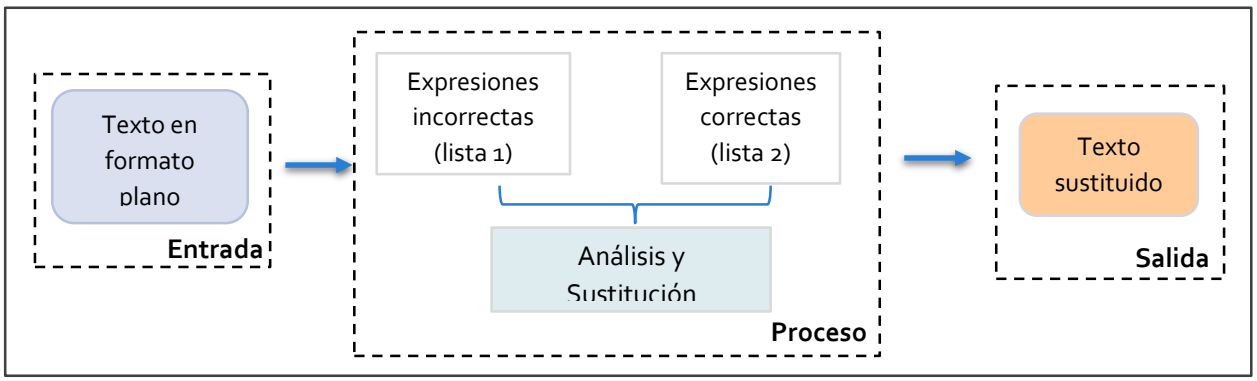


Figura 15 Metodología del sub-modulo “Vicios del lenguaje”

Los datos de entrada a este sub-módulo son en texto plano (txt) y como salida se muestra en consola el texto remplazado correctamente, en la Figura 16 se muestran los resultados de salida de este módulo en consola, se puede observar el párrafo original que se analizó (color rojo) y su correcto reemplazo (color azul).

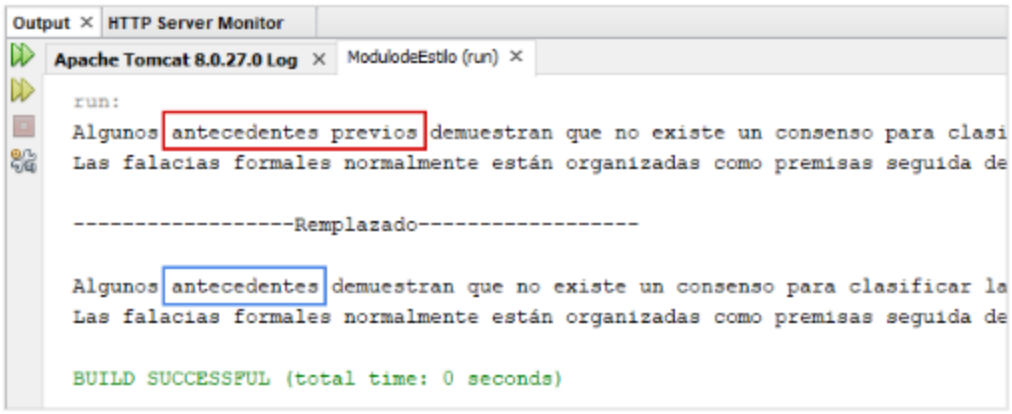


Figura 16 Salida del sub-módulo "Vicios del lenguaje"

- ## Sub-módulo "Uso de la persona gramatical"

Para llevar a cabo este sub-módulo se realizaron dos actividades, en la primera actividad se realizó el etiquetado de las palabras con ayuda de un paquete de herramientas de análisis de lenguaje de código abierto llamado FreeLing [8], el etiquetado se realizó para posteriormente extraer los atributos de las palabras tomando en cuenta los atributos propuestos por el grupo EAGLES [9], el cual propone 12 categorías de etiquetas entre los que se encuentran sustantivos, adjetivos, determinantes, verbos, entre otros. Sin embargo, en este caso la categoría de interés para el desarrollo de este sub-módulo es la categoría de verbos, ya que los verbos en una tesis deben estar en modo impersonal, es decir, en tercera persona por lo tanto los atributos que se deben extraer de la etiqueta de verbos son: "categorías" y "persona" ver Figura 17.

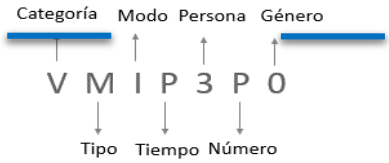


Figura 17 Etiqueta morfosintáctica del verbo

La entrada a este módulo es un archivo en formato txt, el cual contiene tres parámetros (palabra, lema y etiqueta), en este caso se tomó el parámetro 3, como se muestra en la Figura 18.

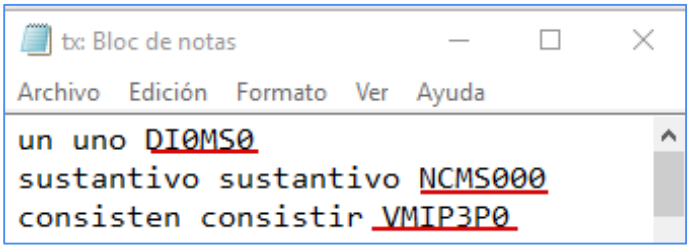


Figura 18 Parámetros de entrada

Para realizar la extracción de los atributos de "categoría" y "persona", se simplificaron las etiquetas utilizando expresiones regulares, sustituyendo los atributos que no son relevantes por el símbolo "." que significa "cualquier carácter", por lo tanto, la siguiente etiqueta < *V M I P 3 P 0* > simplificada quedaría de la siguiente manera: < *V . . . 3 . .* >.

Cabe señalar que el valor del atributo "persona" debe ser siempre "3" ya que este significa que el verbo se encuentra en tercera persona, por lo tanto, si la etiqueta tiene la siguiente estructura < V . . . 1 . . > significa que el verbo está en primera persona, lo cual es incorrecto. Cabe mencionar que las etiquetas de las demás categorías se anulan.

Se realizó un autómata para esta primera actividad, dicho autómata cuenta con dos entradas, una de ellas es para los verbos que están en forma personal <V1>y la otra para los verbos que están en forma impersonal <V3>, en la Figura 19 se muestra el autómata que se desarrolló, en el cual se pueden apreciar las entradas y sus respectivas salidas.

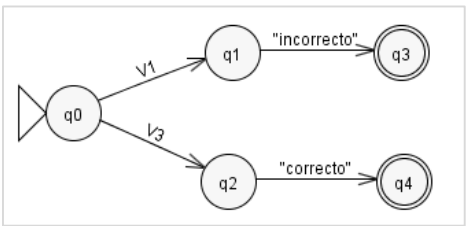


Figura 19 Autómata del sub- módulo "Uso de la persona gramatical"

En la Figura 20 se observa la salida de los datos con una advertencia de correcto (color azul) o incorrecto (color rojo) dependiendo de la etiqueta analizada.

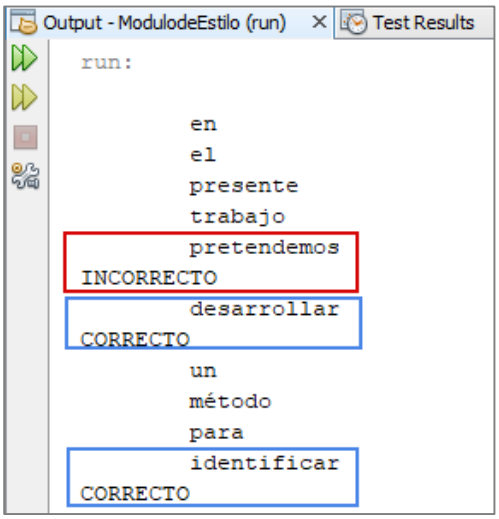


Figura 20 Salida del sub-módulo "Uso de la persona gramatical"

La segunda actividad, se enfoca a la redacción del capítulo "objetivos", para esto también se tomó la categoría de verbos y se simplificaron etiquetas mediante expresiones regulares.

Posteriormente también se creó un autómata de dos entradas, la primera de ellas es para los verbos en infinitivo cuya etiqueta simplificada es <*VMN*>, y la otra para los verbos que están en cualquier otro tiempo verbal. Cabe mencionar que las etiquetas de las demás categorías son eliminadas, es decir, solo se dejan las etiquetas de los verbos.

En la Figura 21 se muestra el autómata que se desarrolló, en el cual se pueden apreciar las entradas y sus respectivas salidas.

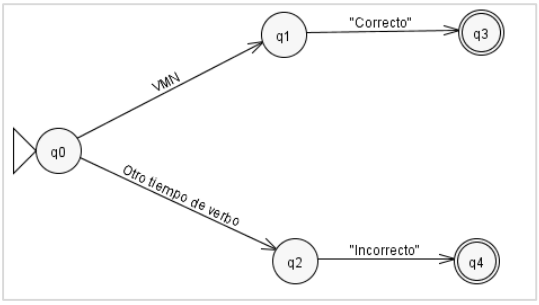


Figura 21 Autómata del sub- módulo "Uso de la persona gramatical"

En la Figura 22 se observa la salida de los datos con una advertencia de correcto (color azul) o incorrecto (color rojo) dependiendo de la etiqueta analizada.

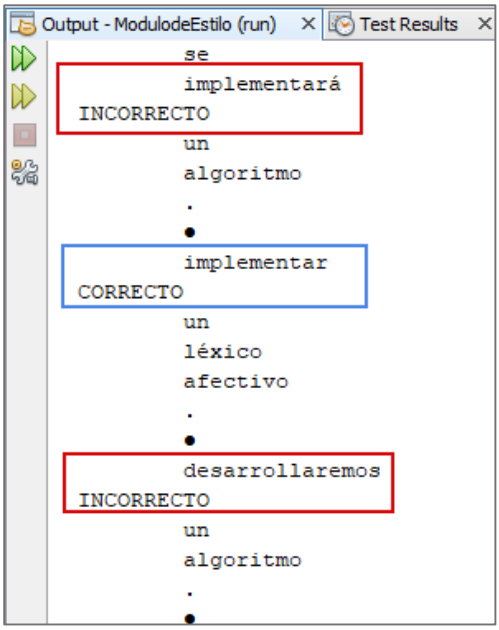


Figura 22 Salida del sub-módulo "Uso de la persona gramatical"

- ## Sub-módulo "Rimas"

Para este sub-módulo se desarrolló un algoritmo que permite eliminar las palabras que cuentan con menos de 4 caracteres, una vez eliminadas las palabras el algoritmo permite segmentar las palabras dejando los últimos tres caracteres, para segmentar las palabras se utilizó el método substring de la clase String el cual permite manipular los caracteres de una cadena.

Después de llevar a cabo la segmentación de las palabras se guardaron en un arreglo para realizar la comparación de las palabras segmentadas, dicha comparación se refleja a continuación.

- Dada la siguiente oración:

  "El diputado no ha disputado lo que se había pronunciado"

- Se realiza la eliminación de palabras con menos de 4 caracteres (color gris)

El	*diputado*	*no*	*ha*	*disputado*	*lo*	*que*	*se*	*había*	*pronunciado*

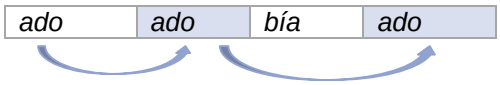
- Quedando de la siguiente manera:

diputado	*disputado*	*había*	*pronunciado*

- Se realiza la segmentación de las palabras con el método substring, quedando de la siguiente manera:

ado	*ado*	*bía*	*ado*

- Posteriormente se realiza la comparación para determinar una igualdad entre los caracteres, tomando en cuenta la distancia entre ellos, por ejemplo:



- Finalmente, si se determina una igualdad entre los caracteres se emite un mensaje de advertencia indicando la existencia de rima en el texto.

La entrada de datos a este sub-módulo está en formato txt, el cual contiene tres parámetros (palabras, lema y etiqueta), el parámetro que se tomó en cuenta es el de “palabras”.

En la siguiente Figura 23 se observa la salida de datos de este sub-módulo.

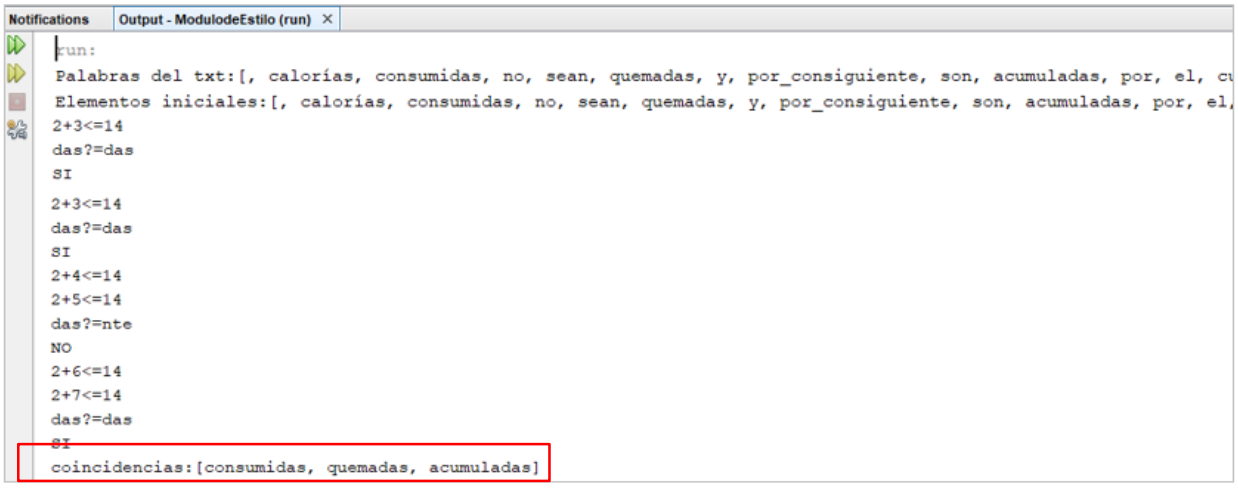


Figura 23 Salida del sub-módulo “rimas”

- **Sub-módulo "Cacofonías"**

En este sub-módulo se desarrolló un algoritmo, el cual permite eliminar las palabras que cuentan con menos de 4 caracteres, después de eliminar esas palabras se lleva a cabo la segmentación de las palabras dejando los primeros tres caracteres, para segmentar las palabras se utilizó el método substring de la clase String.

Una vez realizada la segmentación de las palabras se guardaron en un arreglo para realizar la comparación de las palabras segmentadas, dicha comparación se refleja en el siguiente diagrama.

- Dada la siguiente oración:

  "La clasificación de los clasificadores clásicos"

- Se realiza la eliminación de palabras con menos de 4 caracteres:

La	*clasificación*	*de*	*los*	*clasificadores*	*clásicos*

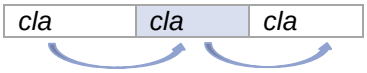
- Quedando de la siguiente manera:

clasificación	*clasificadores*	*clásicos*

- Se realiza la segmentación de las palabras con el método substring, quedando de la siguiente manera:

cla	*cla*	*cla*

- Posteriormente se realiza la comparación para determinar una igualdad entre los caracteres, tomando en cuenta la distancia entre ellos, por ejemplo:



- Finalmente, si se determina una igualdad entre los caracteres se emite un mensaje de advertencia indicando la existencia de rima en el texto.
-

La entrada de datos a este sub-módulo está en formato txt, el cual contiene tres parámetros (palabras, lema y etiqueta), el parámetro que se tomó en cuenta es el de "palabras".

En la siguiente Figura 24 se observa la salida de datos de este sub-módulo.

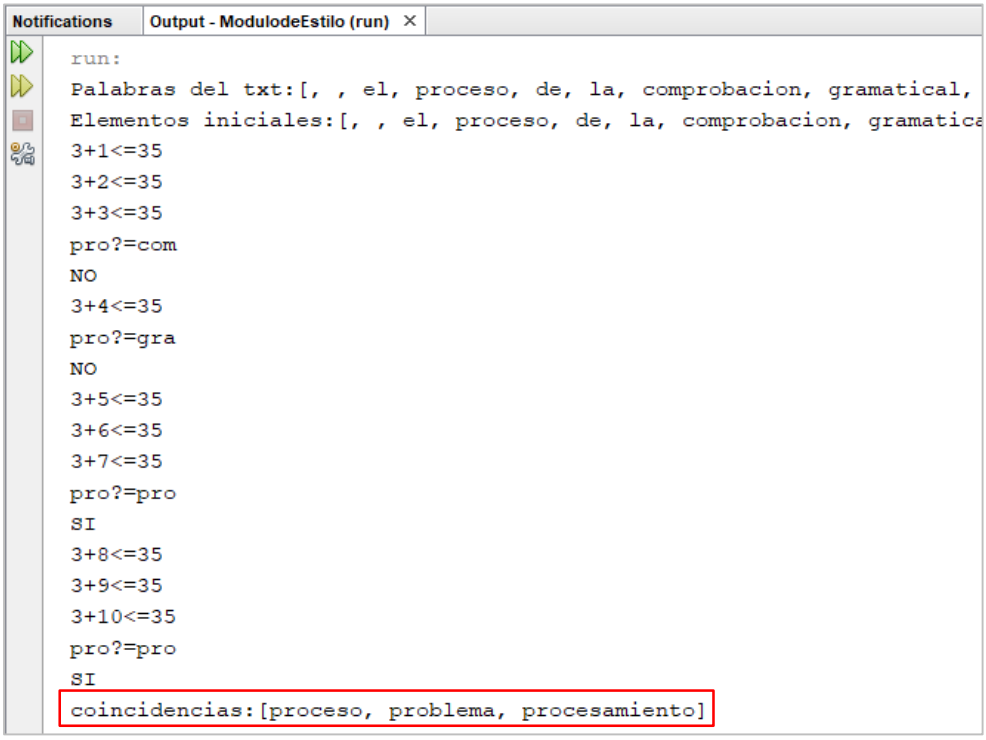


Figura 24 Salida del sub-módulo "cacofonías"

En la Figura 25 se muestra la metodología que se llevó a cabo en el sub-módulo de rimas y cacofonías.

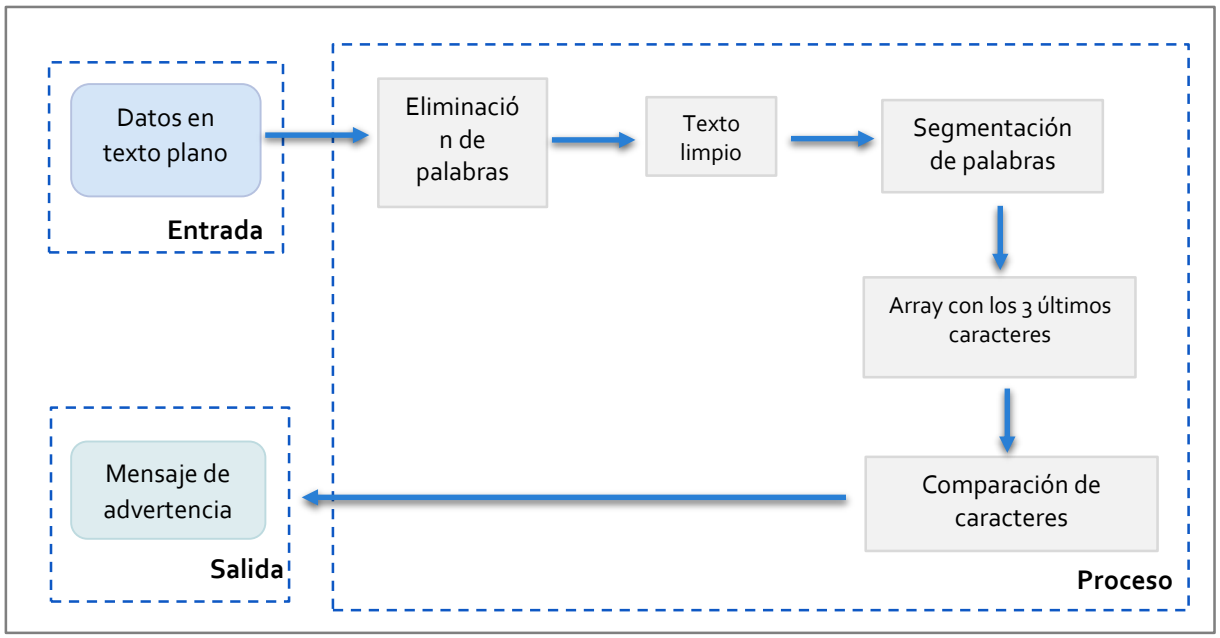


Figura 25 Metodología del sub-módulo "rimas" y del sub-módulo "cacofonías"

- ## Sub-módulo "Jergas populares"

Para esta regla se desarrolló un sub-módulo que consiste en detectar si en el texto se encuentran jergas populares, como recurso para este sub-módulo se cuenta con una lista de jergas populares. La entrada a este módulo es un archivo en formato txt, el cual contiene tres parámetros (palabra, lema y etiqueta), en este caso se toma el primer parámetro "palabras". La metodología empleada en este módulo puede apreciarse en la Figura 26.

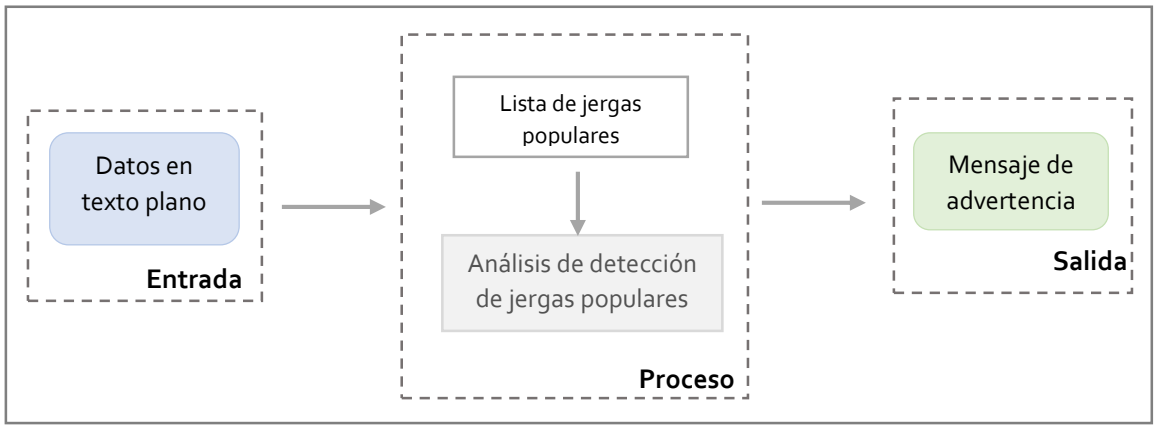


Figura 26 Metodología del sub-módulo "jergas populares"

El análisis consiste en identificar si se encuentran jergas populares en el texto de entrada y si se llega a encontrar una jerga popular se emite un mensaje de advertencia, como se muestra en la Figura 27.

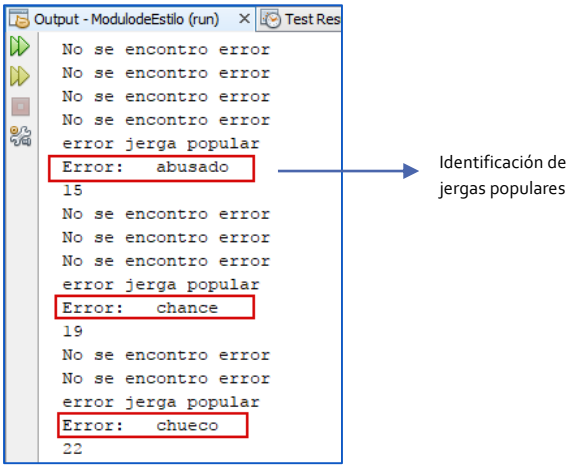


Figura 27 Salida del sub-módulo "jergas populares"

- ## Sub-módulo "Aberraciones lingüísticas"

Para detectar las aberraciones lingüísticas en la tesis se realizó un módulo de programación, el cual cuenta con dos listas, la lista 1 contiene las palabras erróneas que generan deficiencia en la escritura de la tesis y la lista 2 contiene las palabras que pueden sustituirse para evitar este error. La entrada de datos a este módulo es en formato txt, estos datos son analizados y al encontrar alguna expresión que pertenezca a la lista 1 se sustituye por su forma correcta de la lista 2.

Finalmente, como salida se muestra el texto remplazado correctamente, como se aprecia en la Figura 28, el color rojo es la forma incorrecta del uso del lenguaje y en color azul y verde el remplazo correcto.

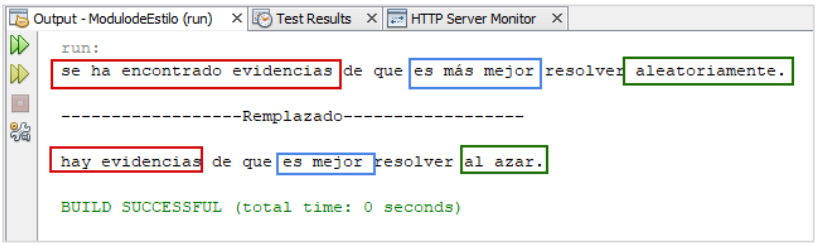


Figura 28 Salida del módulo "aberraciones lingüísticas"

En la Figura 29 se muestra la metodología del módulo aberraciones lingüísticas.

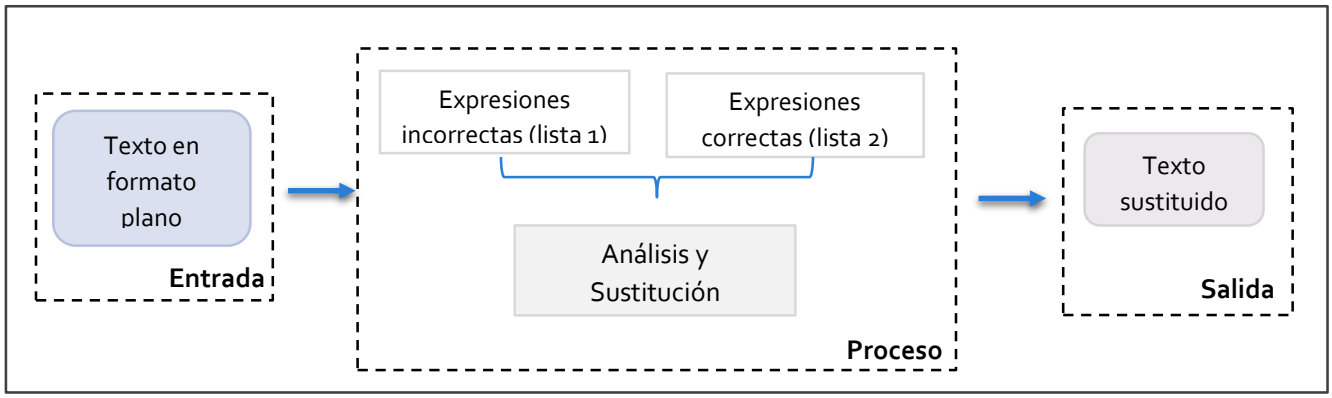


Figura 29 Metodología del módulo "aberraciones lingüísticas"

En la Figura 30 se muestra la metodología con la que se cuenta actualmente para llevar a cabo el análisis de estilo, el cual incluye las reglas mencionadas anteriormente.

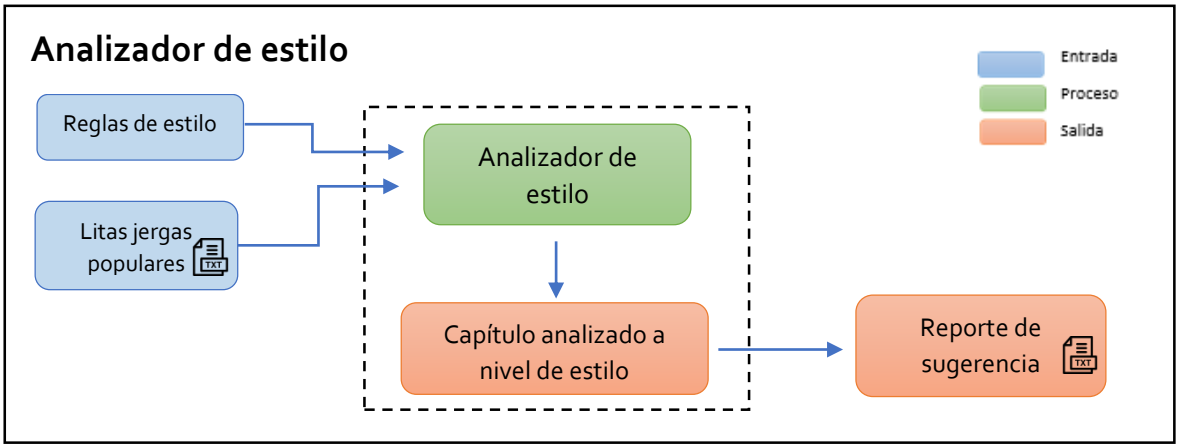


Figura 30 Metodología del módulo "analizador de estilo"

5.2 FASE 2: Análisis de contenido

A continuación se presenta la fase 2, la cual consiste en un análisis de contenido. Esta fase consta de dos etapas, la primera lleva a cabo un análisis de riqueza léxica para proporcionar una mejor calidad de redacción a la tesis y la segunda busca la relación y similitud entre los diferentes capítulos de ésta.

- ### Etapa 1: Analizador de reiteración de vocablos

El siguiente módulo se llama "riqueza léxica", en este módulo se creó un algoritmo que se encarga de analizar el texto y detectar si alguna palabra se reitera varias veces en un párrafo para sugerir algún sinónimo de esa palabra.

La entrada a este módulo, es un texto en formato txt, como primer paso el texto se limpia de aquellas palabras que no aportan ningún significado por sí solas (*stopword*), una vez eliminadas ese tipo de palabras, el texto es fraccionado por párrafos, y cada uno de ellos es almacenado para posteriormente analizar palabra por palabra y detectar igualdad entre ellas, cada igualdad se guarda para ser

comparada con un diccionario de sinónimos, al encontrar igualdad entre las palabras se sugieren los sinónimos que posee esta palabra.

A continuación en la Figura 31 se muestra la metodología usada para lograr la sugerencia de sinónimos.

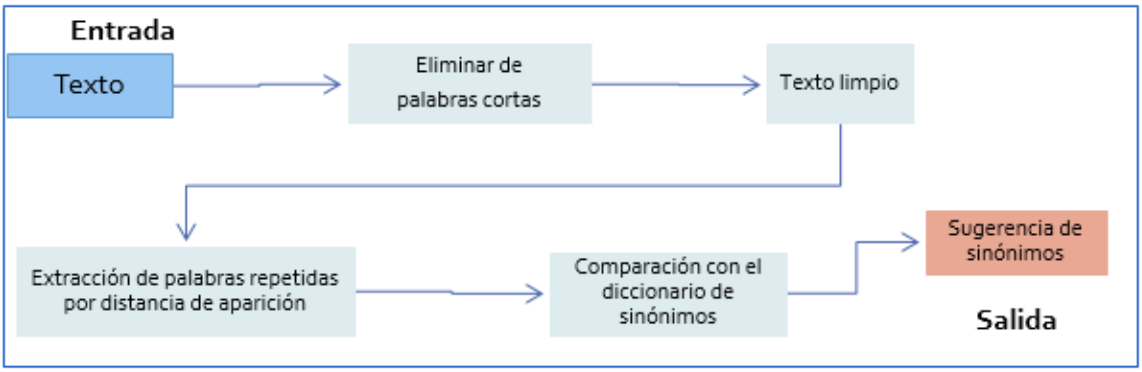


Figura 31 Metodología del módulo "Riqueza léxica"

En la Figura 32 se aprecian los datos de salida de este módulo, se observan las palabras repetidas y las sugerencias de sinónimos de la palabra repetida.

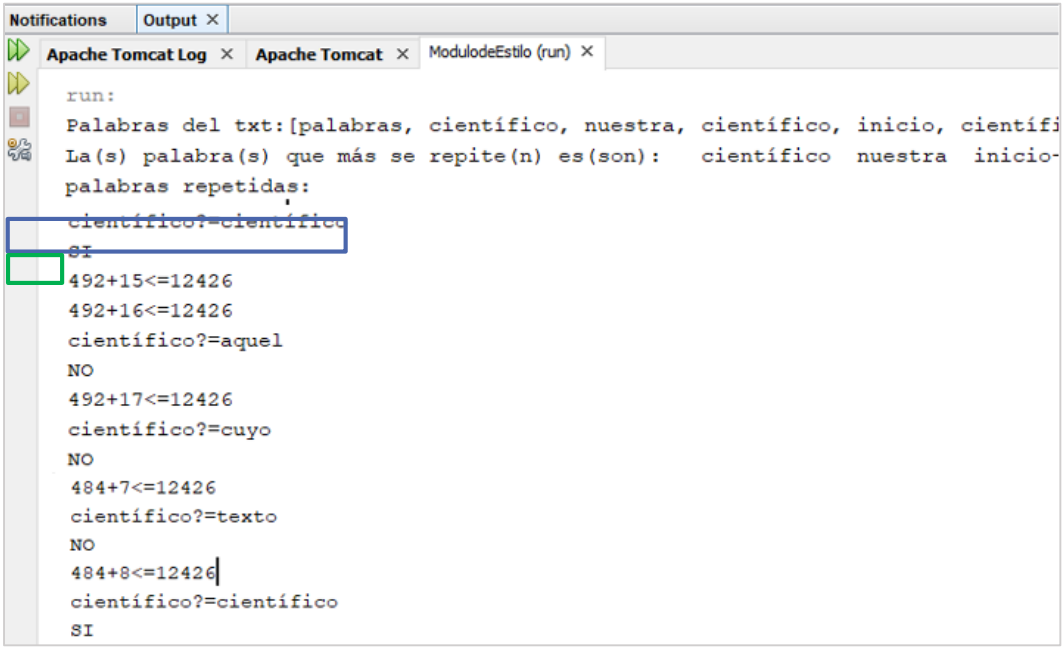


Figura 32 Salida del módulo "repetición de vocablos"

A continuación en la Figura 33 se presenta la metodología empleada en el módulo Analizador riqueza léxica.

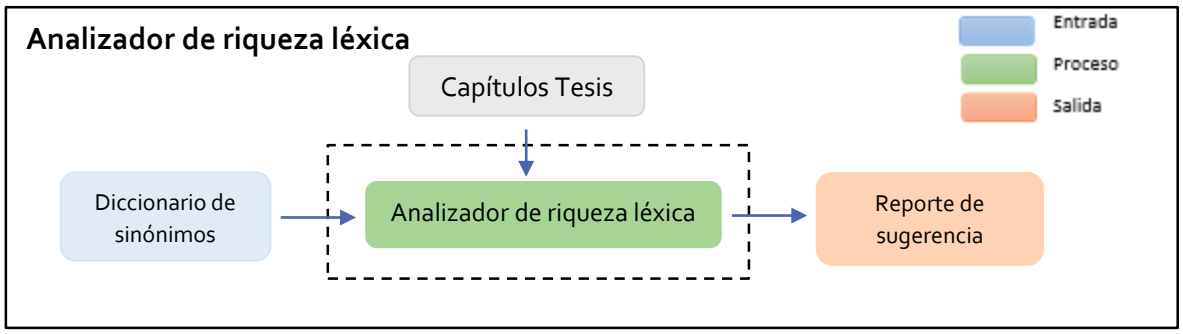


Figura 33 Metodología del módulo Analizador de riqueza léxica

- **Etapa 2: Identificación de similitud semántica entre capítulos**

Esta etapa consistió en la implementación de un algoritmo con el cual se determinó si existe una relación entre los capítulos de la tesis. Los capítulos que se buscan tengan similitud son: "Título" - "Objetivo", "Objetivo" - "Planteamiento del problema" y "Objetivo" - "Conclusión".

Este módulo recibe dos archivos de datos, los cuales están en formato plano (txt), los cuales son analizados mediante el algoritmo "*similitud coseno*", como primer paso, los datos de entrada son preprocesados para eliminar *stopword* y se lematizan las palabras con la finalidad de suprimir aquellas que no tienen importancia y dejando solo las palabras relevantes. Una vez que el texto está libre de *stopword* y lematizado es procesado por el algoritmo de similitud coseno para realizar la comparación de las dos entradas de datos y determinar la similitud existe entre ellos. En el diagrama se observa los pasos que se llevan a cabo para determinar la similitud de los capítulos, ver Figura 34.

Figura 34 Metodología del módulo "Identificación de relación entre capítulos"

En la Figura 35 se muestra la salida que genera este módulo, se puede apreciar el resultado de la comparación de dos textos.

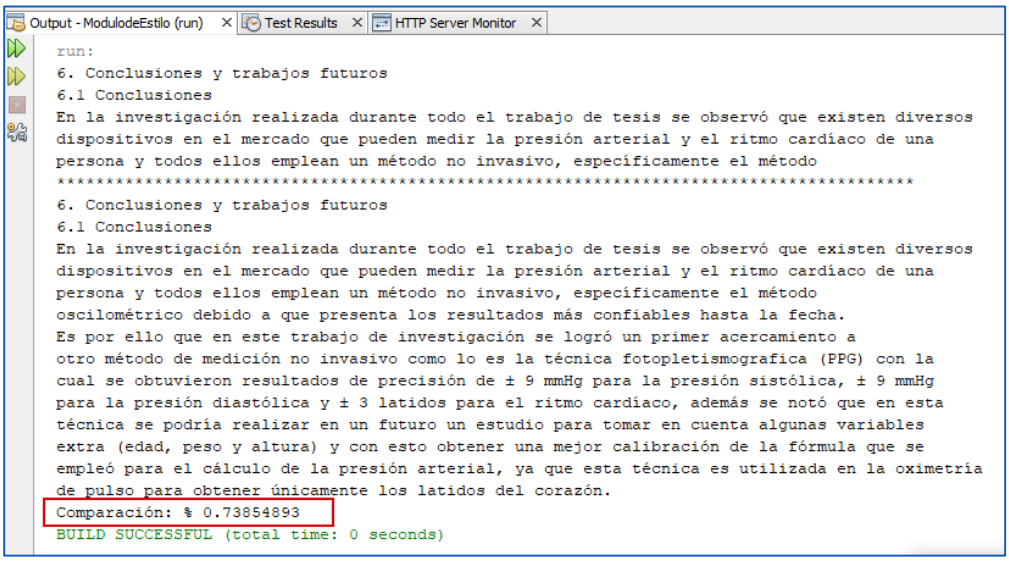


Figura 35 Salida del módulo "Identificación de relación entre capítulos"

A continuación en la Figur4a 36 se muestra la metodología empleada en este módulo.

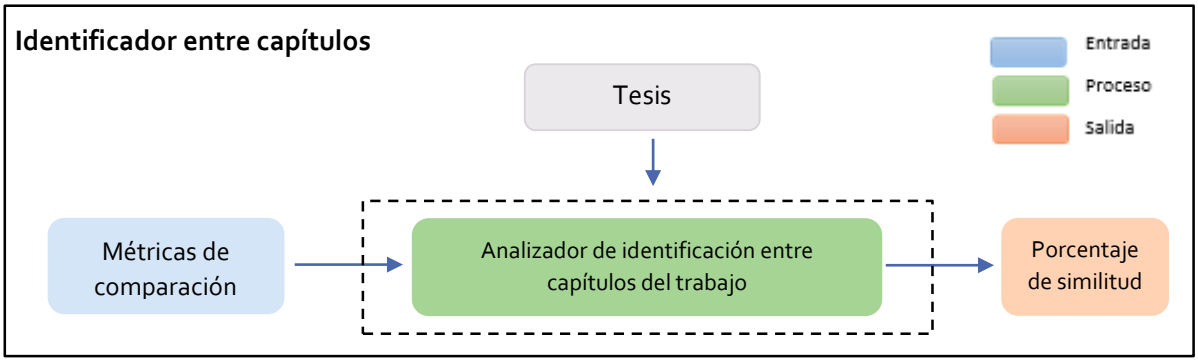


Figura 36 Metodología módulo Identificador entre capítulos

5.3 FASE 3: Generación de reportes

En esta sección se presentan los reportes generados automáticamente de cada módulo de programación, en cada uno de los reportes se adquiere información tal como; el párrafo analizado, el tipo de error detectado, la explicación del tipo de error,

la palabra que presenta el error y por último la posible sugerencia de corrección. Esta información está separada por dos símbolos “#” con la finalidad de ser procesada fácilmente por una computadora. A continuación se presentan los reportes de cada módulo desarrollado.

En la Figura 37 se presenta el reporte de preposiciones.

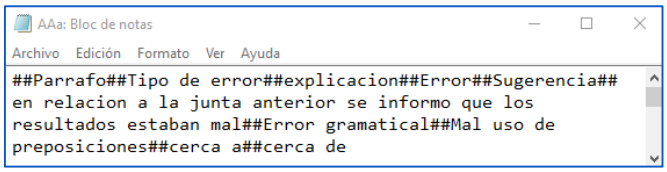


Figura 37 Reporte generado automáticamente del sub-módulo “preposiciones”

En la Figura 38 se observa el reporte de dequeísmo.

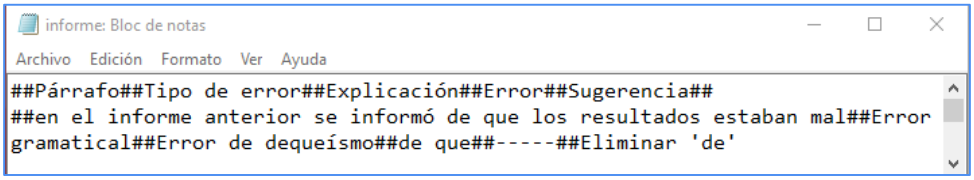


Figura 38 Reporte generado automáticamente del sub-módulo de dequeísmo

En la Figura 39 se muestra el reporte de concordancia nominal y verbal.

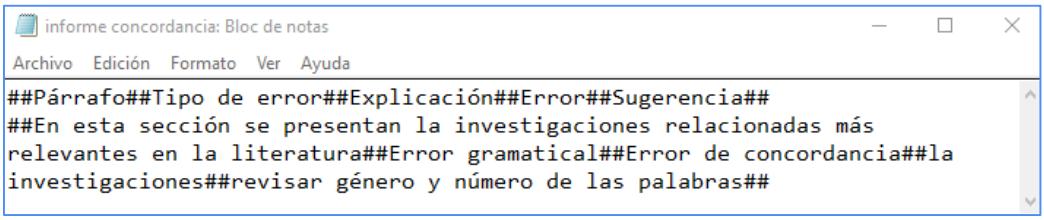


Figura 39 Reporte generado automáticamente del sub-módulo de concordancia nominal y verbal

En la Figura 40 se observa el reporte de palabras tónicas.

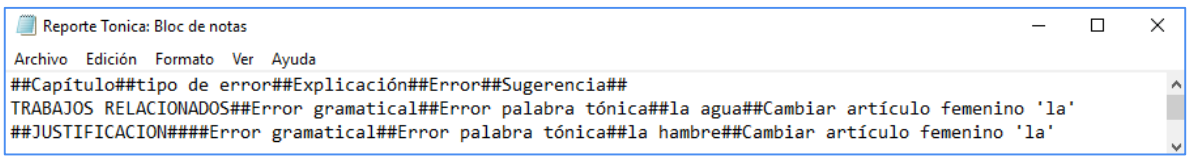


Figura 40 Reporte sub-módulo “Palabras tónicas”

En la Figura 41 se muestra el reporte de expresiones mal empleadas en una tesis.

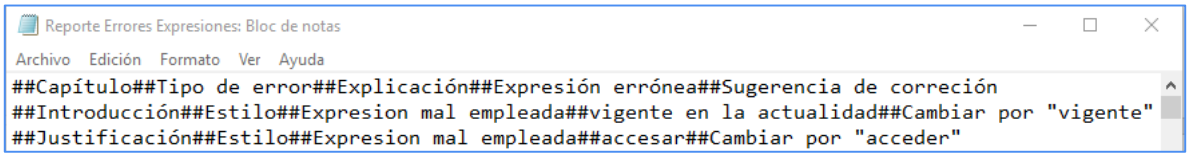


Figura 41 Reporte sub-módulo "vicios del lenguaje"

En la Figura 42 se presenta el reporte del módulo "redacción impersonal".

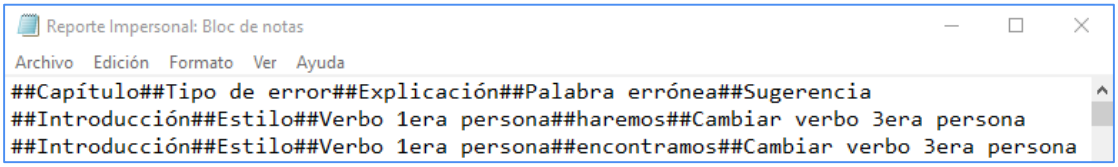


Figura 42 Reporte sub-módulo "Uso de la persona gramatical"

Este reporte, presenta los errores que se generan en el capítulo de objetivos en cuanto al tiempo de los verbos, en la Figura 43 se observa el reporte generado automáticamente.

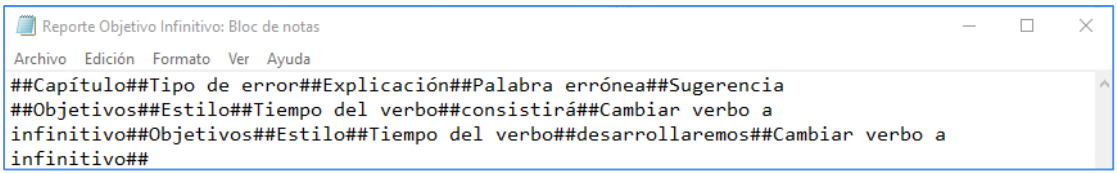


Figura 43 Reporte sub-módulo "Uso de la persona gramatical"

El reporte de rimas y cacofonías se muestra en la Figura 44 y Figura 45 respectivamente.

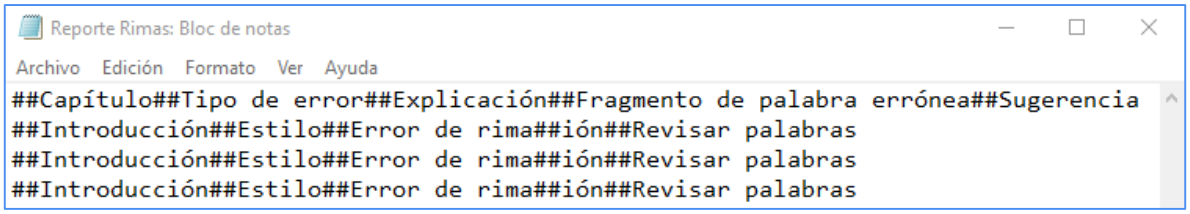


Figura 44 Reporte sub-módulo "rimas"

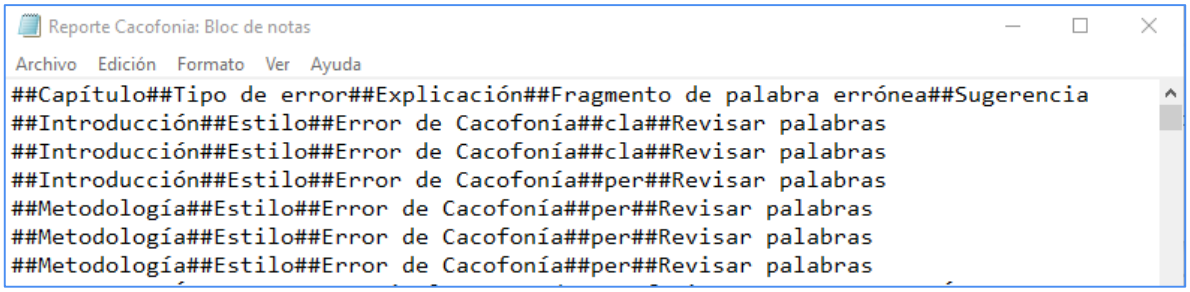


Figura 45 Reporte sub-módulo "cacofonías"

El reporte de jergas populares, en él se muestran los errores encontrados en los capítulos de la tesis, en la Figura 46 se observa el reporte generado automáticamente.

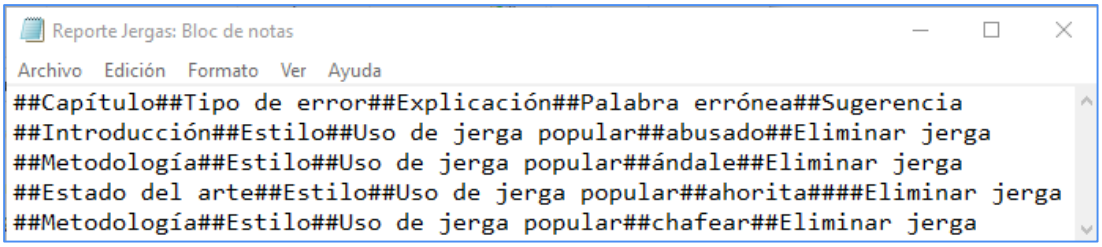


Figura 46 Reporte sub-módulo "jergas populares"

En la Figura 47 se observa el reporte generado automáticamente del sub-módulo de aberraciones lingüísticas.

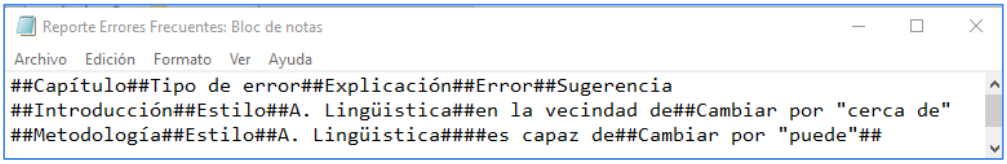


Figura 47 Reporte del módulo "aberraciones lingüísticas"

Capítulo VI

Aplicación Web

Aplicación Web

La aplicación web "Validador de textos científicos" se compone de 2 módulos y cada uno de ellos identifican diferentes tipos de errores, en la Tabla 15 se da una breve descripción de éstos.

Tabla 15 Tipo de errores analizados por el sistema "Validador de textos científicos"

Módulos de la aplicación web "Validador de textos científicos"		
Módulo Gramatical		
Tipo de error	**Descripción**	**Ejemplo**
Dequeísmo	Uso indebido de la preposición "de" delante de la conjunción que cuando la preposición no viene exigida por ninguna palabra en el enunciado.	*"En nuestra vida diaria se dice **de que** los argumentos están presentes también en la información que se trasmite por televisión, la radio, el periódico, personas, internet, etc."*
Preposiciones	Se debe emplear el buen uso de las preposiciones para que cumplan su función de relacionar los elementos de una oración adecuadamente.	*"en relación a" - "en relación con"* *"respecto a" - "con respecto a"* *"en relación a" - "con relación a"*
Preposiciones con verbos	La preposición debe obedecer la noción léxica del núcleo verbal que se encuentra adyacente a ella.	*Lema del verbo + "a" = "atrever a"* *Lema del verbo + "de" = "abusar de"* *Lema del verbo + "en" = "insistir en"*
Concordancia nominal y verbal	La concordancia es la coincidencia obligatoria entre los elementos variables (género, número y persona) de una oración.	*"A **este última** se le conoce como teoría de la argumentación falaz, o falacias…*
Determinante con palabra tónica	Cuando se presenta un sustantivo femenino que comience por /a/ o /ha/ tónica, el artículo femenino "la" que lo acompaña se convierte obligatoriamente en "el" artículo masculino 'el'.	*"La inteligencia artificial (ia) se caracteriza por ser **una área** multidisciplinaria."*
Módulo de estilo		
Tipo de error	**Descripción**	**Ejemplo de errores**
Vicios del lenguaje	Vicios del lenguaje que se encargan de pronunciar o escribir mal algunas palabras, se les conoce como barbarismos, extranjerismos, vulgarismos y pleonasmos.	*"Algunos **antecedentes previos (correcto: antecedentes)** demuestran que no existe un consenso para clasificar las falacias) …"*
Aberraciones lingüísticas	Deficiencias que afectan la claridad del texto científico.	***Se ha encontrado evidencia (correcto: "hay evidencias")** de que en esta disciplina existen diferentes tipos de estudios del discurso."*

Módulo de estilo		
Tipo de error	**Descripción**	**Ejemplo de errores**
Jergas populares	Expresiones que se originan por la edad o los diferentes grupos y sirven para diferenciar un grupo de otro.	*"Este **cacho** de código indica el reemplazo de las oraciones."*
Redacción impersonal	La escritura de una Tesis debe estar en modo impersonal, por lo tanto, se recomienda el uso de la tercera persona.	*"En la gráfica 10 **demostramos** que **obtenemos** mejores resultados, ya que se realiza primero el resumen y por consiguiente la traducción automática. "*
Objetivos en verbos infinitivo	El capítulo de objetivos debe contener verbos en infinitivo al inicio de la oración.	*"**Creación** de un corpus de discursos políticos etiquetado por humanos."*
Cacofonía	Se reconocen como cacofonías a la sucesión de palabras que empiezan con la misma sílaba o el mismo prefijo.	*"El **proceso** de la comprobación gramatical es un **problema** en el **procesamiento** de lenguaje natural (pln), se realizan con el **propósito** de encontrar errores gramaticales..."*
Rimas	La rima se produce cuando existe la sucesión de palabras con igual terminación a poca distancia una de la otra.	*Gran parte de las calorías **consumidas** no sean **quemadas** y por consiguiente son **acumuladas** por el cuerpo.*
Repetición de palabras	Sustitución léxica con la finalidad de mantener el fluido del texto.	*"Uno de los grandes problemas de una **persona** con parálisis cerebral es la interacción con una **persona**, les es muy difícil ya que tienen la limitante de no poder darse a entender de una manera adecuada con las **personas** a su alrededor, en el peor de los casos, ninguna **persona** a su alrededor podría ayudar a la **persona** discapacitada..."*

La aplicación Web analiza errores gramaticales y de estilo y permite subir una tesis en formato .docx, posteriormente hace la segmentación de los capítulos y los almacena en ficheros independientes de tipo.txt. Seguido de esto, cada capítulo es etiquetado automáticamente por la herramienta FreeLing y procesado por la aplicación.

A continuación, en la Figura 48 se muestra la pantalla principal, la cual cuenta con nombre del sistema, la hora y el botón “Seleccionar documento” que permite cargar la Tesis para su procesamiento.

Figura 48 Vista principal de la aplicación Web

En la Figura 49 se observa la pantalla principal del sistema, la cual muestra el análisis de la Tesis, en el marco derecho se observan los tipos de errores que se detectan, así como la cantidad de éstos y un icono de descarga, el cual permite obtener el reporte de los errores detectados.

Figura 49 Tipos de errores analizados por la aplicación Web

En el anexo 1 se muestran ejemplos de la detección de errores de cada módulo.

Capítulo VII

Pruebas

Pruebas

En esta sección, se describen las pruebas que se realizaron en el sistema de "Validador de textos científicos" con la finalidad de confirmar el funcionamiento adecuado de los módulos desarrollados.

Los datos de entrada para realizar las pruebas de los 2 módulos fueron 20 Tesis en idioma español del programa de posgrado de Maestría elaboradas en el Centro Nacional de Investigación y Desarrollo Tecnológico. De dichas tesis se procesaron automáticamente los capítulos de "Introducción", "Objetivos", "Planteamiento del problema", "Resultados" y "Conclusiones", ya que en estos capítulos se exponen los propósitos de la investigación, así como la solución del problema y los resultados obtenidos.

Previo al análisis automático realizado por el sistema, 4 hablantes nativos del idioma español analizaron manualmente la información morfológica de las palabras y sus categorías gramaticales (determinantes, nombres, adjetivos, adverbios y verbos), así como la identificación de errores gramaticales como concordancia nominal, dequeísmo, preposiciones, y por último errores de estilo como jergas populares, redacción en tercera persona, entre otros. Así mismo, un experto en el lenguaje realizó un análisis manual de 4 Tesis para identificar los errores antes mencionados.

A continuación en los siguientes capítulos, se explica el análisis manual y automático, así como las pruebas realizadas de cada uno de los módulos.

7.1 Pruebas Fase 1

- **Pruebas del módulo gramatical**

A continuación se presentan las pruebas que se realizaron para cada uno de los sub-módulos que componen el módulo gramatical, los cuales fueron desarrollados en el lenguaje de programación Java. El módulo gramatical fue puesto a prueba con el corpus de Referencia del Español Actual (CREA) con el propósito de comprobar su correcto funcionamiento. Cabe señalar que existen 2 niveles para realizar las pruebas e identificar los errores, los cuales son: a nivel oración y a nivel palabra.

- ## Pruebas del sub-módulo “Preposiciones”

Para las pruebas con respecto al mal uso de las preposiciones en la redacción de la tesis, se consideraron las oraciones del texto. Este sub-módulo fue puesto a prueba con 20 secciones de “Introducción”, con un total de 6,960 palabras y 1,392 oraciones. A continuación, se muestra un ejemplo de la detección de preposiciones incorrectas.

Texto original con errores de preposiciones

"*Realizar el diseño de una ontología que defina el perfil de usuario* ***en base a*** *las preferencias identificadas y un cuestionario sobre datos personales del usuario.* "

Texto corregido

"*Realizar el diseño de una ontología que defina el perfil de usuario* ***con base en*** *las preferencias identificadas y un cuestionario sobre datos personales del usuario.* "

En la Figura 50 se muestran los resultados de salida de este módulo en consola, se puede observar el párrafo original que se analizó y su correcto reemplazo (color anaranjado).

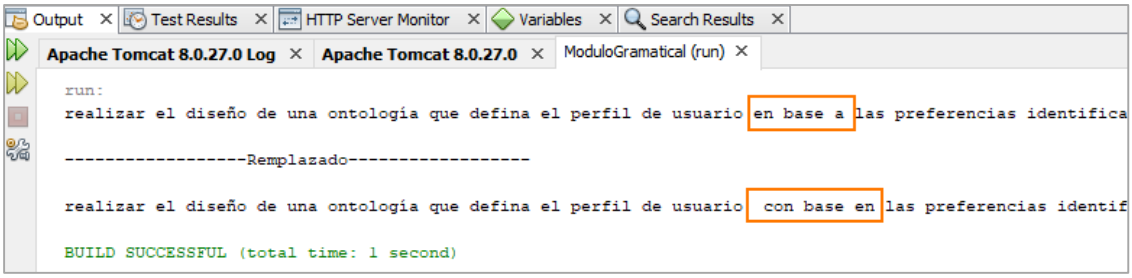


Figura 50 Salida del sub-módulo “Preposiciones”

- ## Pruebas del sub- módulo “Preposiciones con verbos”

En esta sección, se presentan las pruebas que se realizaron con respecto al mal uso de las preposiciones que deben acompañar a ciertos verbos. En este módulo se procesaron 20 secciones de “Introducción”, con un total de 6,960 palabras, de las cuales 912 fueron verbos y de éstos sólo 9 empleaban de forma incorrecta la preposición que los acompaña.

A continuación, se muestra un ejemplo de un texto que contiene preposiciones con verbos de forma incorrecta y su respectiva corrección.

Texto original con errores de preposiciones con verbos:

*"Se han **empezado en** estudiar el tema de las emociones, porque influyen en la conducta y experiencia subjetiva del ser humano. (claves, 2010). En el presente trabajo se **aspira con** desarrollar un método..."*

Texto corregido:

*"Se han **empezado a** estudiar el tema de las emociones, porque influyen en la conducta y experiencia subjetiva del ser humano. (claves, 2010). En el presente trabajo se **aspira a** desarrollar un método..."*

En la Figura 51 se muestra la salida de datos en consola que arroja este módulo, se observa el verbo y la preposición errónea (color verde) y se da un mensaje de sugerencia con la preposición correcta (color azul).

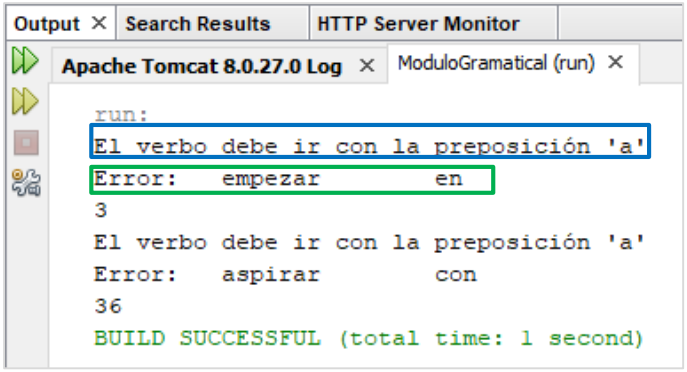


Figura 51 Salida del sub-módulo "Preposiciones con Verbos"

- ## Pruebas del sub-módulo "Dequeísmo"

En seguida se presentan las pruebas del sub- módulo "dequeísmo", para realizarlas se analizaron y procesaron 20 secciones de "Planteamiento del problema", con un total de 7,560 palabras.

A continuación se muestra un ejemplo de la detección de dequeísmo, así como su corrección.

Texto original con errores de dequeísmo:

*"En nuestra vida diaria se dice **de que** los argumentos están presentes también en la información que se trasmite por televisión, la radio, el periódico, personas, internet, etc. también se piensa **de que** nuestras propias ideas sobre algún tema puedan refutar la de vida otros."*

Texto corregido:

> *"En nuestra vida diaria se dice **que** los argumentos están presentes también en la información que se trasmite por televisión, la radio, el periódico, personas, internet, etc. también se piensa **que** nuestras propias ideas sobre algún tema puedan refutar la de vida otros."*

En la Figura 52 se muestra la salida de datos en consola, se observan los verbos con los cuales el dequeísmo no está permitido.

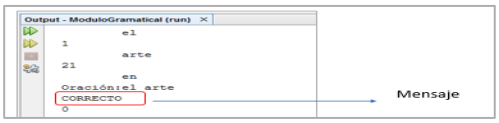


Figura 52 Salida del sub-módulo "Dequeísmo"

- ## Pruebas del sub-módulo "Concordancia nominal y verbal"

A continuación se presentan las pruebas que se realizaron en el sub-módulo de "concordancia nominal y verbal". Para este sub-módulo se realizó un autómata que está alimentado de reglas gramaticales y patrones de concordancia, algunos de ellos son enlistados a continuación en la Tabla 16, el resto de los patrones pueden observarse en el anexo 1, así como su explicación.

Tabla 16 Ejemplos de patrones de concordancia nominal y verbal

#	Patrones
1.	DFS + NFS
2.	DMS + NMS
3.	DFP + NFP
4.	DMP + NMP

El análisis de detección de errores de concordancia se realizó a nivel de oración, ya que se separaron las oraciones del resto del texto para ser analizadas una por una. Posteriormente el sistema fue puesto a prueba con 20 secciones de "Conclusiones", con un total de 4,840 palabras, con un aproximado de 1200 oraciones.

A continuación, se muestra un ejemplo donde se detectan 5 errores de concordancia nominal y verbal.

Texto original con errores de concordancia:

> *"A **este última** se le conoce como teoría de **las argumentación falaz**, o falacias (vega reñón, 2012) a pesar del **crecientes interés** que se está dando a el estudio de **las falacia**, en la actualidad sigue considerándose **un investigación** pobre."*

Texto corregido:

> *"A **esta última** se le conoce como teoría de **la argumentación falaz**, o falacias (vega reñón, 2012) a pesar del **creciente interés** que se está dando a el estudio de **las falacias**, en la actualidad sigue considerándose **una investigación** pobre."*

En la Figura 53 se muestra la salida en consola de una de las oraciones que presenta error de concordancia, se observa la oración errónea (color verde) y un mensaje de "incorrecto" (color rojo).

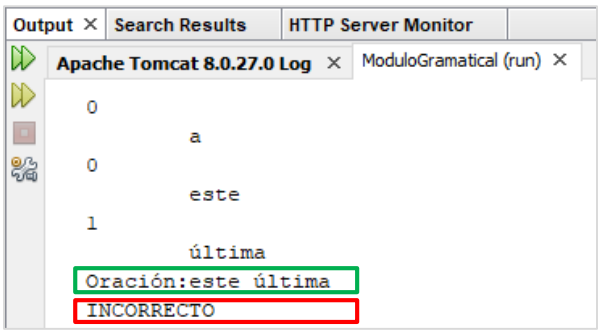


Figura 53 Salida del sub-módulo "Concordancia nominal y verbal"

- ## Pruebas del sub-módulo "Anáfora"

En esta sección, se muestra un análisis de una de las reglas de anáforas, la cual consiste en analizar la coincidencia entre el "*sustantivo colectivo + preposición "de" + verbo*". El análisis de detección de errores de anáfora se realizó a nivel de oración, ya que también se separaron las oraciones del resto del texto para ser analizadas una por una. Para realizar las pruebas se analizaron 20 secciones de "Conclusiones", con un total de 4,840 palabras con un aproximado de 1,200 oraciones. A continuación se muestra un ejemplo de la detección de errores de anáfora.

Texto original con errores de anáfora:

> "*el equipo de futbolistas **están** cansado*"

Texto corregido:

> "*el equipo de futbolistas **está** cansado*"

En la Figura 54 se muestra la salida de datos en consola, como se puede apreciar, la oración analizada "*el equipo de futbolistas están…*" presenta discordancia entre el sujeto y el verbo, por lo tanto, se muestra un mensaje de incorrecto (color rojo).

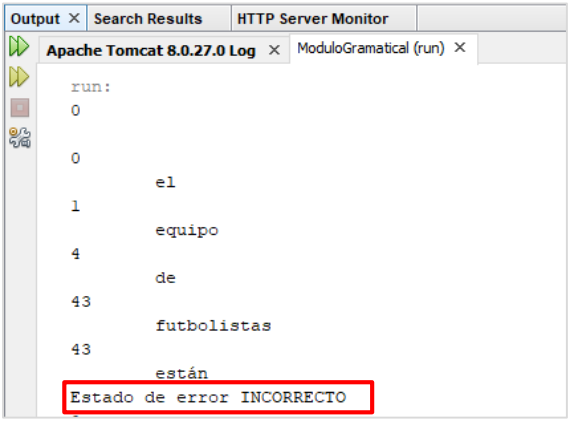


Figura 54 Salida del sub-módulo "Anáforas"

- ## Pruebas del sub-módulo "Palabras tónicas"

En este sub-módulo también se realizó un autómata el cual analiza la regla "*determinante masculino + sustantivos con 'a' y 'ha' tónica*". En este caso, para realizar las pruebas se analizaron y procesaron 20 Tesis completas, con un total de 517,720 palabras.

A continuación se muestra un ejemplo de la detección de errores con palabras tónicas.

Texto original con errores con palabras tónicas:

> "*La inteligencia artificial (ia) se caracteriza por ser **una área** multidisciplinaria.*"

Texto corregido:

> "*La inteligencia artificial (ia) se caracteriza por ser **un área** multidisciplinaria.*"

En la Figura 55 se muestra la salida de datos en consola, en donde se aprecia el error (color rojo) y un mensaje de sugerencia (color azul).

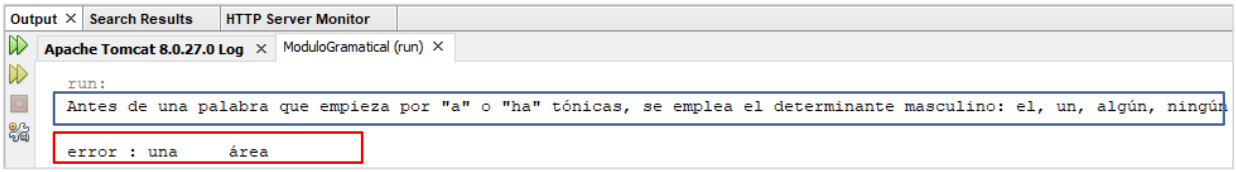


Figura 55 Salida del sub-módulo "Tónica"

- ## Pruebas del módulo de estilo

En esta sección se presentan las pruebas que se realizaron para cada uno de los sub-módulos que componen el módulo de estilo, los cuales fueron desarrollados en el lenguaje de programación Java. El módulo de estilo fue puesto a prueba con el corpus de Referencia del Español Actual (CREA) con la finalidad de comprobar su correcto funcionamiento. Cabe señalar que existen 2 niveles para realizar las pruebas e identificar los errores, los cuales son: a nivel oración y a nivel palabra.

- ## Pruebas del sub-módulo "Vicios del lenguaje"

En esta sección, se presentan las pruebas que se realizaron en el sub-módulo de "vicios del lenguaje". El análisis de detección de errores se realizó a nivel de palabras y fue puesto a prueba con 20 secciones de "Conclusiones", con un total de 4,840 palabras. A continuación, se muestra un texto, el cual presenta vicios del lenguaje.

Texto original con errores

"*Algunos **antecedentes previos** demuestran que no existe un consenso para clasificar las falacias, **de otra parte**, debido a su naturaleza compleja la manera en que son clasificadas no concuerda entre los expertos del tema.* "

Texto corregido:

"*Algunos **antecedentes** demuestran que no existe un consenso para clasificar las falacias, **por otra parte**, debido a su naturaleza compleja la manera en que son clasificadas no concuerda entre los expertos del tema.* "

En la Figura 56 se muestran los resultados de salida de este módulo en consola, se puede observar el párrafo original que se analizó (color rojo) y su correcto reemplazo (color azul).

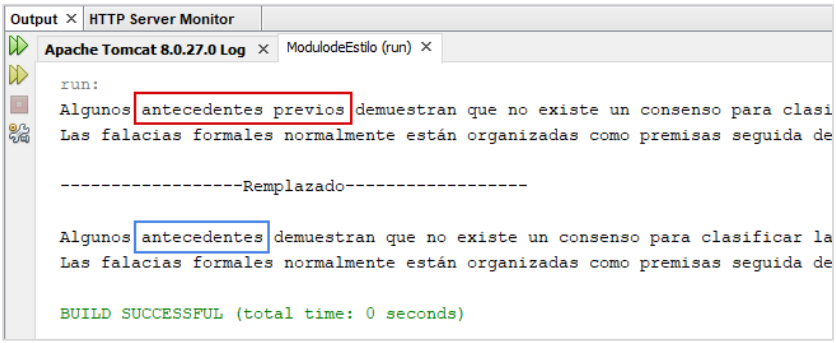


Figura 56 Salida del sub-módulo "Vicios del lenguaje"

- ## Pruebas del sub-módulo "Uso de la persona gramatical"

Para las pruebas de este sub-módulo se analizaron y procesaron 20 secciones de "Introducción", con un total de 6,960 palabras, de las cuales 912 fueron verbos. Dichos verbos fueron analizados manual y automáticamente para determinar cuáles de ellos se encuentran en forma impersonal y cuáles en forma personal.

A continuación, se muestra un ejemplo de redacción personal y su correcta redacción para un texto científico.

Texto original redactado en forma personal

> *"En la gráfica 10 **demostramos** que **obtenemos** mejores resultados, ya que se realiza primero el resumen y por consiguiente la traducción automática. "*

Texto corregido:

> *"En la gráfica 10 se **demuestra** que se **obtienen** los mejores resultados ya que se realiza primero el resumen y por consiguiente la traducción automática. "*

En la Figura 57 se muestran los resultados de salida de este sub-módulo en consola, se observan los verbos analizados y se determina cuáles de éstos están en forma impersonal (color azul) y cuales se encuentran en forma personal (color rojo). También se muestra la cantidad de dichos verbos en la Figura 58.

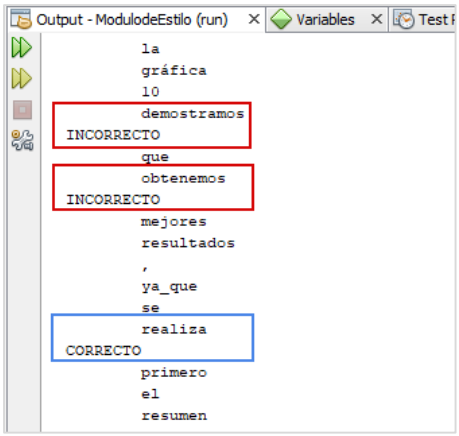


Figura 57 Salida del sub-módulo "Uso de la persona gramatical"

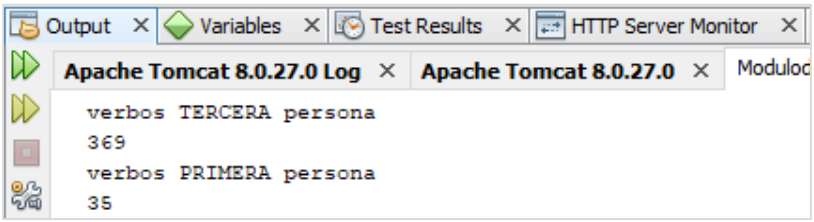


Figura 58 Salida del sub-módulo "Uso de la persona gramatical"

Por otro lado, el capítulo de "Objetivos" debe estar redactado con verbos en infinitivo al inicio de cada uno de los objetivos. Los datos de entrada a este sub-módulo fueron 20 secciones de "Objetivos", con un total de 2,940 palabras, de los cuales 235 son verbos, sin embargo, solo se realizaron pruebas con los verbos que se encuentran al inicio de la oración, por consiguiente, sólo se analizaron 120 verbos en total.

A continuación, se muestra un ejemplo de redacción de objetivos de forma correcta e incorrectamente.

Texto original redactado de forma incorrecta:

"***Creación*** *de un corpus de discursos políticos etiquetado por humanos."*

Texto corregido:

"***Crear*** *un corpus de discursos políticos etiquetado por humanos."*

En la Figura 59 se muestran los resultados de salida de este módulo en consola, se puede apreciar el análisis del primer elemento de la oración "*creación*" y como no es un verbo en infinitivo el sistema lo identifica como incorrecto (color rojo).

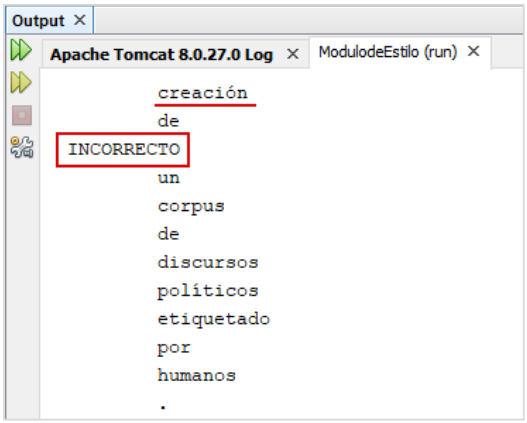


Figura 59 Salida del sub-módulo "Uso de la persona gramatical"

- ## Pruebas del sub-módulo "Rimas"

Antes de presentar las pruebas, se debe mencionar que las rimas presentes en la tesis son accidentes gramaticales. No obstante, la rima intencionada es usada como una expresión poética o como una actividad lúdica, es por ello que se realizó un estudio para determinar las características entre una rima intencionada y una causada por error del escritor. A continuación, se presentan dos versos que contienen rimas y se pueden apreciar sus características:

*Inmóvil en la luz, pero danz**ante,***

*tu movimiento a la quietud que cr**ía***

*en la cima del vértigo se al**ía*** → Distancia de 7 palabras

Distancia de 21 palabras ← *deteniendo, no al vuelo, sí al inst**ante.***

*Luz que no se derrama, ya diam**ante,***

*fija en la rotación del mediod**ía,***

*sol que no se consume ni se enfr**ía*** → Distancia de 8 palabras

Distancia de 19 palabras ← *de cenizas y llama equidist**ante.***

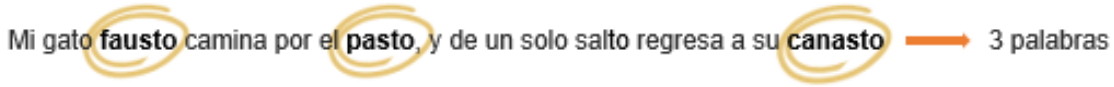
En los ejemplos anteriores se muestran las rimas de un poema, pero también existen rimas pequeñas (con menor distancia entre las palabras con la misma terminación), las cuales se analizan a continuación en la Tabla 17 , en dicha tabla se muestran el número de letras, el número de palabras y por último, la distancia entre una palabra y otra con la misma terminación.

Tabla 17 Análisis de rimas

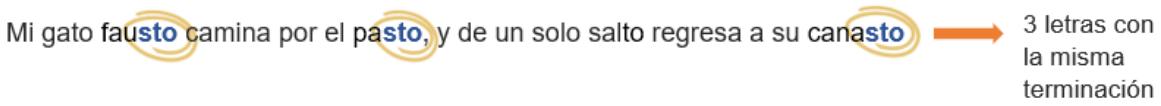
Ejemplos	# de letras	# de palabras	Rango de distancia entre palabras	Presencia de rimas
• A la víb*ora*, víb*ora* de la m*ar*, por aquí pueden pas*ar*… • A las tr*es*, bajito *es*. • Beat*riz* está fel*iz*, como una lomb*riz*.	2	2 2 3	4 2 2-3	Sí Sí Sí
• Fabul*oso*, un limpiador fragante y olor*oso* • A la *una*, sale la L*una*. • Mi muñeca se llama Sabr*ina*, es muy linda y cam*ina*. • Mi gato fau*sto* camina por el pa*sto*, y de un solo salto regresa a su cana*sto*.	3	2 2 2 3	5 3 5 4-9	Sí Sí Sí Sí
• A las c*inco*, doy un br*inco*. • De rubio cab*ello*, mi muñeca es un dest*ello*.	4	2 2	3 5	Sí Sí

En los ejemplos anteriores, se muestra que para que exista una rima deben existir en la oración de 2 -3 palabras que tengan de 2 – 4 letras con la misma terminación y el rango entre dichas palabras sea de 2-9, a continuación se presenta un ejemplo de lo dicho anteriormente.

Número de palabras en la oración:

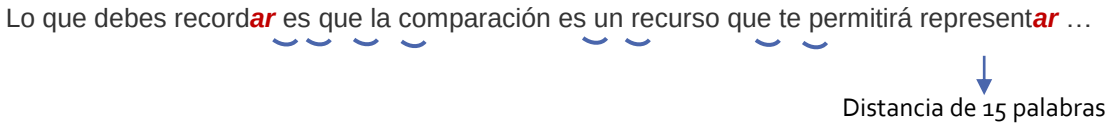


Número de letras en las palabras:



Distancia entre las palabras:

Ahora bien, en el siguiente ejemplo se muestra un texto, el cual presenta 2 palabras con la misma terminación pero la distancia entre las palabras es de 15, es decir, supera el rango de 2 a 9, por lo tanto no se presenta una rima, dicho ejemplo refuerza los parámetros tomados en cuenta en la Tabla 18.



Sin embargo, estas características de rimas no son similares a las que se presentes en una tesis, es por ello que se realizó un estudio y se demuestra que para exista una rima deben existir en la oración más de 2 palabras que tengan 3 letras con la misma terminación y la distancia entre las palabras deben de estar en un rango de 1-8.

Lo anterior se puede comprobar con los ejemplos de la Tabla 18, en la cual se demuestra que aunque existan palabras con la misma terminación, no se genera una rima al menos que en la oración haya mínimo 3 palabras con la misma terminación.

Tabla 18 Análisis de rimas en tesis

Ejemplos	# de letras	# de palabras	Rango de distancia entre palabras	Presencia de rimas
• Hemos recogido tod*as* las apariciones en el corpus de sintagm*as* • A través de est*os* criteri*os* de búsqueda, hem*os*… • pero algun*os* otr*os* cas*os* son exclusiv*os* de ese…	2	2 2 4	7 1-3 1-2	No No No
• una rupt*ura* total con la concepción romántica de la literat*ura* • esta separac*ión* lleva a la consider*ación* de… • como consecuencia de la potenciac*ión* que sufre la ensoñac*ión*, la confus*ión* puede… • en el que se almacenan imág*enes*, sensacio*nes* e impresio*nes*	3	2 2 3 3	8 4 2-4 1-2	Sí Sí Sí Sí

Dados los ejemplos anteriores se puede concluir que las rimas presentes en las tesis deben tener más de 2 palabras que tengan mínimo 3 letras con la misma terminación para que suene repetitivo y el rango de distancia entre las palabras es 1-8.

El análisis de detección de errores de este módulo se realizó a nivel de palabra y para llevar a cabo las pruebas, se creó un corpus de más de 400 oraciones con rimas, éstas fueron descargadas de libros digitales de poesía española contemporánea y de portales web dedicados a la rima en textos. Estas oraciones fueron validadas por un experto con el objetivo de comprobar que el sistema identifique correctamente este tipo de error.

Sin embargo, como ya se había mencionado, las rimas en las tesis tienen diferentes características, por ello se analizaron 20 secciones de "Resultados", con un total de 5,920 palabras, de donde se obtuvo que el texto presentaba la cantidad de 960 palabras con rimas, este análisis fue validado también por el experto en el uso del lenguaje.

A continuación se muestra un ejemplo de la detección de errores de rimas en la redacción de Tesis.

Texto original, el cual presenta rimas:

> *"Gran parte de las calorías consumi**das** no sean quema**das** y por consiguiente son acumula**das** por el cuerpo. "*

En la Figura 60 se muestran los resultados de salida de este módulo en consola, se puede apreciar el análisis de la terminación de las palabras y en color rojo se señalan las palabras con terminaciones iguales, las cuales generan errores de rimas.

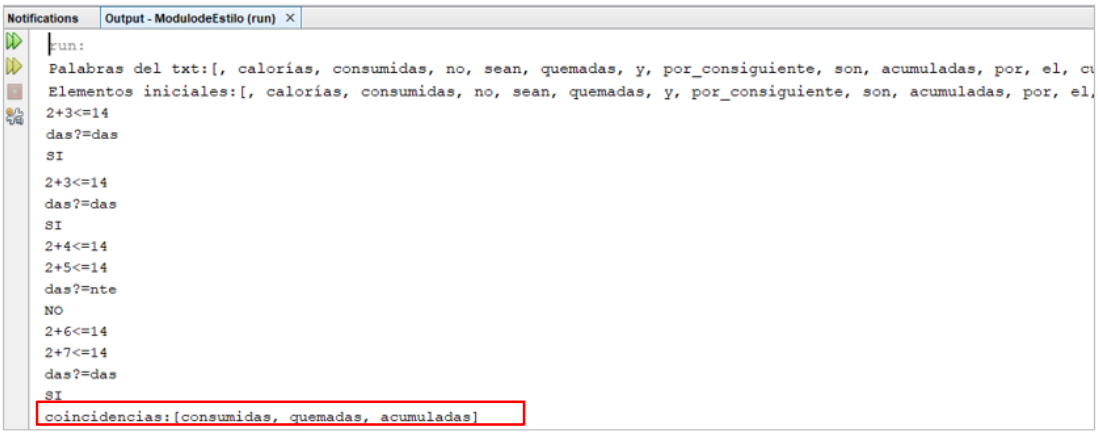


Figura 60 Salida del sub-módulo "rimas"

- **Pruebas del sub-módulo "Cacofonías"**

Al igual que las rimas, las cacofonías que se presentan en las tesis son accidentes gramaticales; sin embargo, el uso de éstas en otro tipo de textos puede utilizarse en varias ocasiones como un recurso literario o para generar efectos humorísticos, tal es el caso de los trabalenguas, es por esta razón que se realizó un estudio de las características entre las cacofonías intencionadas y las que son generadas por error del escritor. En la Tabla 19 se observan ejemplos de cacofonías intencionadas y sus características, como el número de letras, el número de palabras y la distancia entre una palabra y otra con la misma pronunciación.

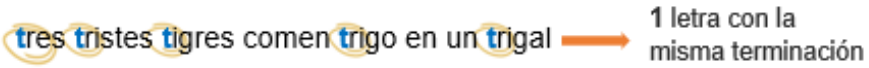
Tabla 19 Análisis de cacofonías

Ejemplos	# de letras	# de palabras	Rango de distancia entre palabras	Presencia de cacofonías
• yo amo a **mi** **m**amá, **mi** **m**amá ama a **m**imo	1	5	1 -2	Sí
• **pí**a fuma **pi**pa • no **co**mas **co**co **co**n los **co**dos • **tr**es **tr**istes tigres comen **tr**igo en un **tr**igal	2	2 4 5	2 1-2 1-3	Sí Sí Sí
• **Mar**ta vendrá **mar**eada el **mar**tes. • **cál**mate tomándote un **cal**mante.	3	3 2	1-2 3	Sí Sí
• Pablo **clav**ó un **clav**ito, ¿dónde **clav**ó el **clav**ito Pablito?	4	4	1-2	Sí

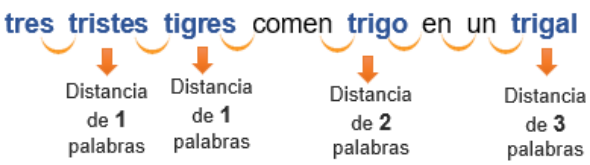
En los ejemplos anteriores, se demuestra que para que exista una cacofonía deben existir en la oración de 2 - 5 palabras que tengan mínimo 1 letra igual y el rango entre las palabras es de 1 - 4.

Número de palabras en la oración:

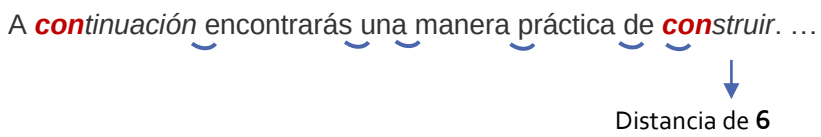
Número de letras en las palabras:



Distancia entre las palabras:



Por otro lado, en el siguiente ejemplo se muestra un texto el cual presenta 2 palabras que tienen 3 letras con la misma pronunciación pero la distancia de las palabras es de 6; es decir, supera el rango de 1 a 4, por lo tanto no se presenta cacofonía. Dicho ejemplo refuerza los parámetros tomados en cuenta en la Tabla 19.



Por otra parte, en las tesis la cacofonía no es intencionada, más bien son accidentes gramaticales que cometen los estudiantes (tesistas) al momento de redactar. Por esta razón, se realizaron estudios de este tipo de error en las tesis y se demuestra que deben existir en la oración más de 2 palabras con la misma pronunciación y la distancia entre las palabras que presentan cacofonía debe tener un rango de 1-6.

Lo anterior se puede comprobar con los ejemplos de la Tabla 20, en la cual se demuestra que aunque existan palabras con la misma pronunciación, no se genera una cacofonía al menos que en la oración haya mínimo 3 palabras con la misma pronunciación.

Tabla 20 Análisis de cacofonías en tesis

Ejemplos	# de letras	# de palabras	Rango de distancia entre palabras	Presencia de Cacofonías
• se constituye como un caso extremo	1	3	1 -2	No
• en nuestra vida diaria se dice	2	2	2	No
• un caso extremo de la mayor cantidad de atributos		2	5	No
• se han empezado en estudiar el tema de las emociones		2	7	No
• los diccionarios fluctúan en su distancia semántica		2	4	No
• a este última se le conoce como teoría de …		2	1	No
• en esta disciplina existen diferentes tipos de estudios del discurso		3	1-5	No
• de este modo se estableció la correspondencia	3	2	2	No
• esta organización y estudio conforman…		2	2	No
• se realiza la concordancia con un número incorrecto		2	1	No
• como paso previo a la presentación de los resultados		2	3	No
• diseñaremos y construiremos un dispositivo		2	3	No
• las restricciones de la participación son problemas para participar en situaciones vitales		3	1-3	No
• el proceso de la comprobación gramatical es un problema en el procesamiento		3	2-6	Sí

Dados los anteriores ejemplos se puede concluir que las cacofonías presentes en las tesis, presentan mayor distancia entre las palabras (hasta de 6) que generan

una cacofonía y también el número de palabras es mayor a 2, las cuales deben contener mínimo 3 letras iguales al inicio de la palabra.

Para efectuar las pruebas, se creó un corpus de más de 60 oraciones que presentan cacofonías, las cuales fueron descargadas de portales web dedicados a la lengua y la literatura; estas oraciones fueron validadas por un experto con el objetivo de comprobar que el sistema identifique correctamente este tipo de error en el texto. Sin embargo, como se mencionó, las cacofonías presentes en las tesis tienen diferentes características, por esta razón se analizaron también 20 secciones de "Conclusiones" con un total de 4,840 palabras, de dicho análisis se obtuvo que el texto presentaba la cantidad de 540 palabras con cacofonía, este análisis fue validado también por un experto en el uso del lenguaje. A continuación, se muestra un ejemplo de la detección de errores de cacofonías en la redacción de Tesis.

Texto original, el cual presenta cacofonías:

> *"El **pro**ceso de la comprobación gramatical es un **pro**blema en el **pro**cesamiento de lenguaje natural (pln), se realizan con el **pro**pósito de encontrar errores gramaticales en un texto de entrada. "*

En la Figura 61 se muestran los resultados de salida de este módulo en consola, se puede apreciar el análisis de los primeros caracteres de las palabras, las cuales son señaladas en color rojo.

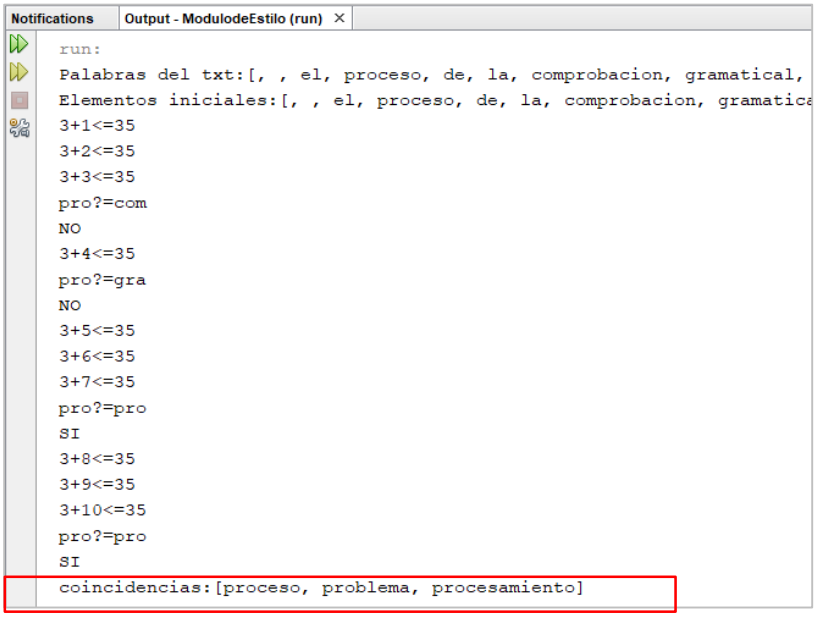


Figura 61 Salida del sub-módulo "rimas"

- **Pruebas del sub-módulo "Jergas populares"**

El análisis de detección de errores de jergas populares se realizó a nivel de palabra, y fue puesto a prueba con 20 secciones de "Conclusiones", con un total de 4,840 palabras, de las cuales sólo 5 de ellas fueron identificadas como errores de jergas populares. A continuación, se muestra un ejemplo de la detección de errores de jergas populares.

Texto original, el cual presenta jergas populares

"Se obtuvo un par de chances y se escogió el más complicado…"

En la Figura 62 se muestra el análisis de este módulo, el cual consiste en identificar si se encuentran jergas populares en el texto y si se llega a encontrar una de ellas se emite un mensaje de error (color rojo).

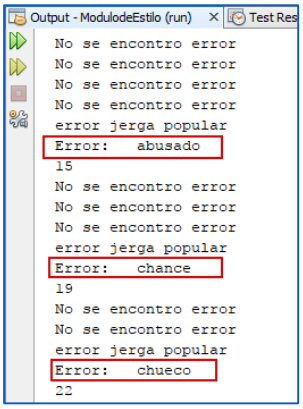


Figura 62 Salida del sub-módulo "jergas populares"

- **Pruebas del sub-módulo "Aberraciones lingüísticas"**

Para las pruebas de este sub-módulo se analizaron 20 secciones de "Planteamiento del problema", con un total de 7,560 palabras. El análisis se llevó a cabo de forma manual y automática, los resultados obtenidos de dicho análisis reportan 18 aberraciones lingüísticas en los capítulos analizados. A continuación, se muestra un ejemplo de la detección de errores de jergas populares y su corrección.

Texto original:

> *"**Durante el transcurso** de 60 años se dio inicio (particularmente en Francia) una serie de investigaciones sobre el discurso político; estas investigaciones dieron origen a una nueva disciplina llamada análisis del discurso (Murillo Serna, 2001). **Se ha encontrado evidencia** de que en esta*

Texto remplazado:

> *"**Durante** 60 años se dio inicio (particularmente en Francia) una serie de investigaciones sobre el discurso político; estas investigaciones dieron origen a una nueva disciplina llamada análisis del discurso (Murillo Serna, 2001). **Hay evidencia** de que en esta disciplina existen diferentes tipos de estudios del discurso*

En la Figura 63 se muestran los resultados de salida de este módulo en consola, se pueden observar lo errores del párrafo original que se analizó en color rojo y su correcto reemplazo en color azul.

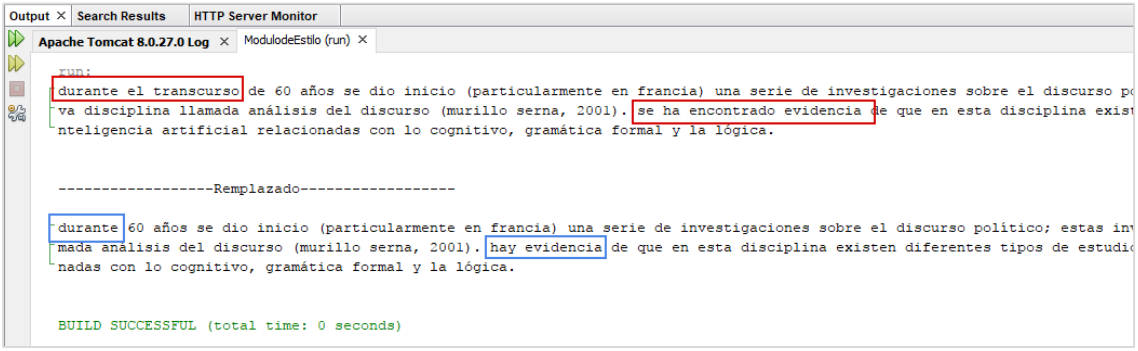


Figura 63 Salida del módulo "aberraciones lingüísticas"

7.2 Pruebas Fase 2

- **Pruebas del sub-módulo "Repetición de vocablos"**

Para las pruebas de este sub-módulo se analizaron 10 párrafos de diferentes tesis por 4 hablantes nativos del idioma español, los cuales identificaron aquellas palabras que presentaban repetición. Es necesario recalcar que se tomaron en consideración las palabras que se agrupan en las categorías gramaticales de sustantivo, verbos, adjetivos y adverbios, debido a que otras categorías como preposiciones, conjunciones, pronombres y artículos son de uso frecuente y generan una repetición inevitable y que según Marouzeau[6], son llamadas "palabras

[6] Lingüista y filólogo francés nacido en Fleural (Creuse) en 1878 y muerto en Iteuil (Vienne) en 1964.

accesorias". En la Figura 64 se ilustra el análisis manual llevado a cabo por los hablantes nativos del idioma español.

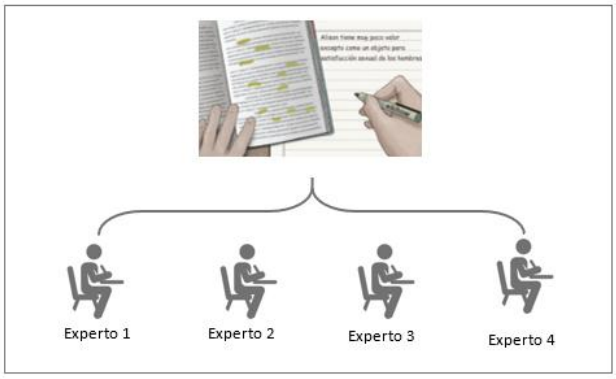


Figura 64 Análisis manual de repetición de vocablos

Como ya se mencionó, para el análisis manual se les presentaron a los hablantes nativos del idioma español diversos párrafos extraídos de tesis del programa de posgrado de Maestría del CENIDET, en donde existe repetición de palabras con proximidad cercana y lejana.

Posteriormente, se analizó si los expertos habían coincidido en señalar las palabras que consideraban repetidas (los resultados de este análisis se encuentran en el capítulo 0), así mismo, un experto en el uso del lenguaje analizó 4 tesis e identificó las palabras que se repetían en los párrafos.

Además del análisis manual se llevó a cabo un análisis automático, el cual constó del procesamiento de 20 secciones de "Introducción", con un total de 6,960 palabras, en donde se identificó mediante el análisis que se repetían 420 palabras. En el siguiente ejemplo se muestra un párrafo, el cual fue procesado por el sistema y los datos de salida se observan en el siguiente párrafo.

> *"es por ello que los científicos, para efectuar la comunicación científica se utiliza el lenguaje científico, que posteriormente será divulgado mediante textos científicos con el fin de propagar nuevos conocimientos científicos en una disciplina."*

En la Figura 65 se muestran los datos de salida de este módulo, en donde pueden apreciarse las palabras repetidas. Conviene subrayar que el sistema puede identificar tanto las palabras "raíz" como sus derivaciones, por ejemplo: "científicos"

"científica", ya que el sistema analiza el lema de las palabras. Otro rasgo que se observa, es que se señalan las palabras analizadas en color azul y en color verde se muestra si hay o no repetición.

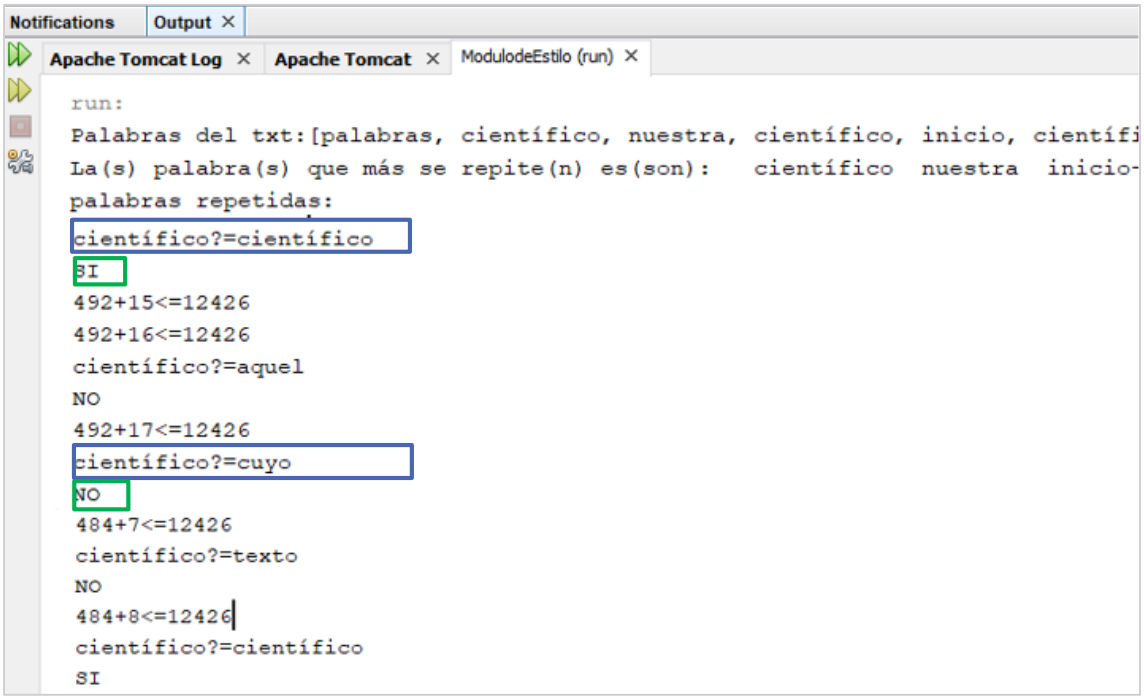


Figura 65 Salida del módulo "repetición de vocablos"

Es importante mencionar que la diferencia de opinión de los expertos en cuanto a que, si una palabra es repetida o no, se debe a que para ellos algunas palabras deben repetirse necesariamente y no pueden omitirse, esto se refuerza con lo que mencionó Pascal cuando dijo: "*Cuando en un discurso encontramos palabras repetidas y, al intentar la corrección, nos damos cuenta de que, al corregir, estropearíamos el discurso, hay que dejar tales palabras*".

A continuación en la Tabla 21 se muestra el análisis manual realizado por los hablantes nativos, los cuales indican la distancia entre las palabras que ellos consideraban era repetida.

Donde:

- *ID párrafo*: se refiere al número de identificación del párrafo que está siendo analizado.
- *ID Experto*: se refiera al número de identificación del hablante nativo del idioma español que analiza el texto

- *Distancia*: se refiere a la distancia entre las palabras que se repiten, en el siguiente ejemplo se muestra la forma en que los hablantes nativos establecieron la distancia de las palabras:

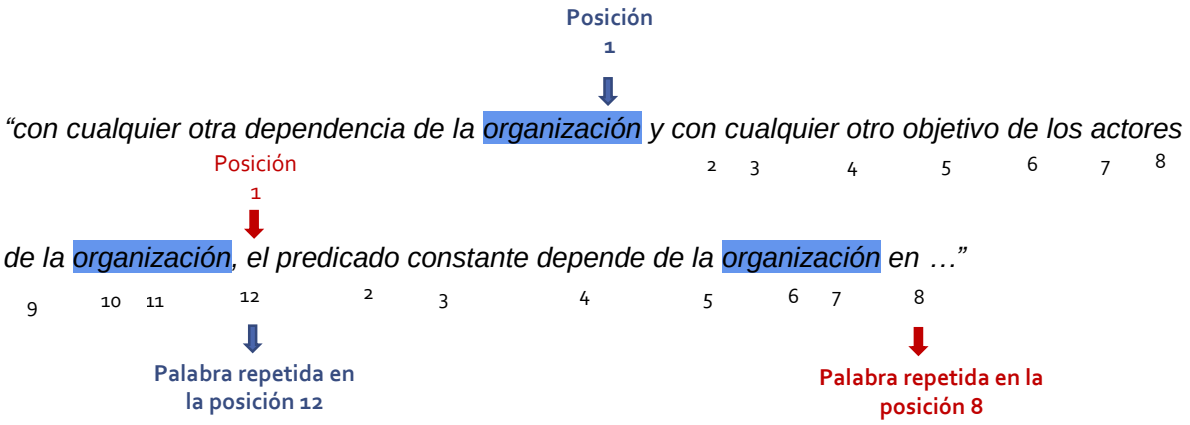


Tabla 21 Análisis manual de repetición de vocablos realizado por los hablantes nativos

ID párrafo	Distancia			
	ID experto			
	ID 1	ID 2	ID 3	ID 4
1	3 7 14	3 7 14	3 7 14	3 7 0
2	2 5	2 5	2 5	2 0
3	3 4 7 9	3 4 7 0	3 4 0 0	3 4 7 9
4	2 13	2 0	2 13	2 13
5	2 6 17	2 6 0	2 6 17	2 6 17
6	13	0	13	13
7	3 11 13	3 11 0	3 0 0	3 11 13
8	2 4 19	2 4 0	2 4 0	2 4 19
9	4 5	0 0	4 0	4 5
10	2 3	2 0	0 0	2 3

Como se mencionó anteriormente, los resultados de las pruebas realizadas pueden verse a detalle en el capítulo de Resultados.

- ## Pruebas del módulo "Identificación de similitud semántica entre capítulos"

En la elaboración de tesis existen dos requisitos para lograr su comprensión y elegancia. El primero de ellos se define como forma que se refiere al uso apropiado del lenguaje y a la organización del texto, mientras que el segundo trata del fondo, el cual abarca la armonía de todas las ideas, la demostración del análisis que debe conducir a las conclusiones, la profundidad que trata de la esencia del problema y la originalidad, la cual se logra mediante el análisis de los intentos realizados por otros investigadores o por el propio investigador de resolver problemas.

Así bien, con la finalidad de obtener una mejor calidad en la tesis se recomiendan queden demostradas las relaciones entre los siguientes capítulos[7]:

- "Título" - "Objetivo"
- "Objetivo" - "Planteamiento del problema"
- "Objetivo" - "Conclusión".

En la Tabla 22, se muestran estas relaciones, así como la explicación de su importancia.

Tabla 22 Comparación de capítulos de una tesis

Relación	Importancia
"Título – Objetivo"	El título representa la esencia de la investigación de forma concreta y generalmente se define brevemente el objetivo de ésta.

[7] http://www.upv.es/laboluz/master/seminario/textos/umberto_eco.pdf

Relación	Importancia
“Objetivo” - “Planteamiento del problema”	El planteamiento del problema debe contener los argumentos que describan el interés de la investigación a realizar, contiene información acerca los planteamientos específicos que identifican lo que se desea lograr con el proyecto de investigación. También se describe la finalidad de la investigación y en ésta se hace referencia al objetivo general y objetivos específicos los cuales corresponden al enfoque propuesto. Es decir, los objetivos responden con exactitud a la definición del problema.
“Objetivo-Conclusión”	La conclusión es la argumentación fundamentada de la problemática y en ésta se dan a conocer los resultados obtenidos. También, se da a conocer si se cumplieron los objetivos planteados.

Cada una de las relaciones (combinaciones) de los capítulos fue analizada manualmente por 5 hablantes nativos del idioma español para determinar si existe similitud entre ellos. Una vez teniendo esos resultados, las combinaciones también se analizaron automáticamente por el algoritmo de similitud coseno, el cual devuelve un valor que indica el porcentaje de similitud que existe entre un par de textos.

Cabe señalar, que el objetivo del análisis manual y automático es determinar el rango de los valores en los que se agrupan los porcentajes, para determinar a partir de qué valor se considera que existe similitud entre los capítulos.

En las siguientes tablas se pueden observar las comparaciones manuales y automáticas que se llevaron a cabo, en las cuales se pueden observar los siguientes parámetros:

- ID de comparación: Se refiere al número de identificación de la comparación entre capítulos que está siendo analizada.
- Similitud encontrada por hablantes: Este campo está dividido en:
    - ID evaluador: Se refiera al número de identificación del hablante nativo del idioma español que analiza el texto, dichos hablantes marcan con un “Sí” si existe similitud entre los capítulos y un “NO” si no existe.
- Mayoría: Este campo se refiere a la mayoría de “SÍ” o “NO” que se encuentran en la columna “Similitud encontrada por hablantes”.

- Porcentaje de similitud del algoritmo: Se refiere al valor que arroja el algoritmo de similitud coseno.

A continuación en la Tabla 23, se muestra las 10 comparaciones entre los capítulos de “Título” y “Objetivos”, se pueden apreciar los valores del análisis manual y el automático.

Tabla 23 Comparación entre Titulo y Objetivos

	Comparación “Título - Objetivo”						
ID de comparación	Similitud encontrada por hablantes					Mayoría (DECISIÓN EXCLUYENTE)	Porcentaje de similitud del algoritmo
	ID evaluador						
	ID 1	ID 2	ID 3	ID 4	ID 5		
1	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.622 %
2	NO	NO	NO	NO	NO	**NO**	0.246 %
3	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.611 %
4	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.609 %
5	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.693 %
6	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.631 %
7	SÍ	NO	NO	NO	NO	**NO**	0.445 %
8	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.653 %
9	SÍ	NO	SÍ	SÍ	SÍ	**SÍ**	0.694 %
10	SÍ	NO	NO	NO	NO	**NO**	0.497 %

En la Tabla 24 se muestra las 10 comparaciones entre los capítulos de “Planteamiento del problema” y “Objetivos”, se pueden apreciar los valores del análisis manual y el automático.

Tabla 24 Comparación entre Planteamiento del problema y Objetivos

	Comparación “Planteamiento del problema - Objetivo”						
ID de comparación	Similitud encontrada por humano					Mayoría (DECISIÓN EXCLUYENTE)	Porcentaje de similitud del algoritmo
	ID evaluador						
	ID 1	ID 2	ID 3	ID 4	ID 5		
01	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.841 %
02	NO	NO	NO	SÍ	NO	**NO**	0.510 %
03	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.584 %
04	SÍ	SÍ	SÍ	NO	SÍ	**SÍ**	0.741 %
05	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.658 %
06	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.769 %
07	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.854 %
08	NO	NO	NO	NO	NO	**NO**	0.488 %
09	SÍ	NO	SÍ	SÍ	SÍ	**SÍ**	0.656 %
010	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.840 %

En la Tabla 25 se muestra las 10 comparaciones entre los capítulos de “Objetivos” y “Conclusión”, se pueden apreciar los valores del análisis manual y el automático.

Tabla 25 Comparación entre Objetivos y Conclusión

ID de comparación	Comparación “Objetivo - Conclusión”						
	Similitud encontrada por humano					Mayoría (DECISIÓN EXCLUYENTE)	Porcentaje de similitud del algoritmo
	ID evaluador						
	ID 1	ID 2	ID 3	ID 4	ID 5		
001	SÍ	SÍ	SÍ	NO	SÍ	**SÍ**	0.877 %
002	NO	NO	NO	NO	SÍ	**NO**	0.528 %
003	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.559 %
004	NO	NO	NO	NO	NO	**NO**	0.465 %
005	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.724 %
006	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.942 %
007	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.766 %
008	NO	NO	NO	NO	NO	**NO**	0.422 %
009	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.734 %
0010	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**	0.751 %

Como ya se mencionó, el parámetro “Mayoría” se determinó dependiendo de la frecuencia en que aparecen en las columnas de “Similitud encontrada por hablantes”, a continuación en la Figura 66 se muestra un ejemplo para mayor comprensión:

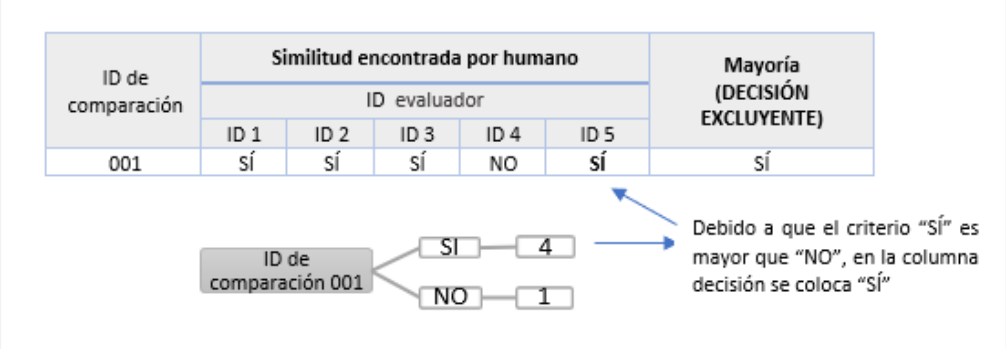


Figura 66 Determinación del parámetro “Mayoría”

En total se realizaron 150 comparaciones. Dado que las opiniones de los evaluadores diferían, se utilizó el índice Kappa de Fleiss[8], el cual mide el grado de concordancia de las evaluaciones realizadas por múltiples evaluadores.

[8] https://www.fisterra.com/mbe/investiga/kappa/kappa2.pdf

A continuación en la Tabla 26 se muestran los datos organizados para el análisis de concordancia de atributos.

Donde:

- **# de comparación:** Se refiere al número de comparación analizada por los hablantes nativos (evaluadores).
- **ID comparación:** Contiene el número de identificación de la comparación entre capítulos.
- **Respuesta:** Es el criterio que decidieron los evaluadores, "SÍ" si consideran que hay similitud y "NO" si no existe. Se encuentra dividida en:
    - **ID Evaluador:** Contiene el identificador del hablante nativo del idioma español, va desde el ID 1 al ID 5.
- **Estándar:** Es el criterio definitivo de similitud.

Tabla 26 Datos para la evaluación de concordancia entre las comparaciones

# de comparación	ID comparación	Respuesta					Estándar
		ID Evaluador					
		ID 1	ID 2	ID 3	ID 4	ID 5	
1	1	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
2	2	NO	NO	NO	NO	NO	**NO**
3	**3**	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
4	4	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
5	5	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
6	6	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
7	7	SÍ	NO	NO	NO	NO	**NO**
8	8	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
9	9	SÍ	NO	SÍ	SÍ	SÍ	**SÍ**
10	10	SÍ	NO	NO	NO	NO	**NO**
11	01	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
12	02	NO	NO	NO	SÍ	NO	**NO**
13	03	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
14	04	SÍ	SÍ	SÍ	NO	SÍ	**SÍ**
15	05	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
16	06	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
17	07	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
18	08	NO	NO	NO	NO	NO	**NO**
19	09	SÍ	NO	SÍ	SÍ	SÍ	**SÍ**
20	010	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
21	001	SÍ	SÍ	SÍ	NO	SÍ	**SÍ**
22	002	NO	NO	NO	NO	SÍ	**NO**

# de comparación	ID comparación	Respuesta ID Evaluador					Estándar
		ID 1	**ID 2**	**ID 3**	**ID 4**	**ID 5**	
23	003	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
24	004	NO	NO	NO	NO	NO	**NO**
25	005	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
26	006	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
27	007	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
28	008	NO	NO	NO	NO	NO	**NO**
29	009	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**
30	0010	SÍ	SÍ	SÍ	SÍ	SÍ	**SÍ**

Para determinar el análisis de concordancia de atributos por evaluador se utilizó la siguiente fórmula[9]:

$$100 \times \frac{número\ de\ evaluaciones\ que\ coinciden\ con\ el\ valor\ estándar\ para\ el\ i^{ésimo}\ evaluador}{N_i}$$

Donde:

- N_i es el número de evaluaciones para el evaluador^{iésimo}

Aplicando las fórmulas a los datos de la Tabla 26, se obtienen los resultados de la que pueden apreciarse en la Tabla 27. Donde:

- **ID Evaluador:** Contiene el identificador del hablante nativo del idioma español.
- **# de comparaciones** Se refiere al número de comparaciones.
- **# de coincidencias:** Se refiere al número de coincidencia que tuvo cada evaluador con relación al total de comparaciones.
- **Porcentaje:** Es la tasa de coincidencia por cada evaluador.
- **IC (nivel de confianza):** Es el grado de certeza (o probabilidad), expresado en porcentaje con el que se quiere realizar la estimación de un parámetro a través de un estadístico muestral, en este caso es del 95 %.

Tabla 27 Acuerdo de evaluación por evaluador

Acuerdo de evaluación				
Evaluador	**# de comparaciones**	**# de coincidencias**	**Porcentaje**	**IC de 95%**
ID 1	30	28	93.33 %	(77.93, 99.18)
ID 2	30	28	93.33 %	(77.93, 99.18)
ID 3	30	30	100.00 %	(90.50, 100.00)
ID 4	30	27	90.00 %	(73.47, 97.89)
ID 5	30	29	96.67 %	(82.78, 99.92)

[9] https://support.minitab.com/es-mx/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/supporting-topics/attribute-agreement-analysis/kappa-statistics-and-kendall-s-coefficients/

Como se puede apreciar en la Tabla 27, cada evaluador clasificó 30 comparaciones, por ejemplo: el evaluador con "ID 1" tuvo 28 coincidencias del total de elementos inspeccionados (comparaciones). También puede apreciarse que todos los evaluadores tienen tasas de coincidencia adecuadas, desde el "ID 1" con 93.33% hasta el "ID 4" con un 90.00%.

Ahora bien, para determinar el análisis de concordancia de atributos total, se utilizó la siguiente fórmula[10]:

$$100 \times \frac{X}{N}$$

Donde:

- X es el número de evaluaciones que coinciden con el valor estándar
- N es el número de filas de datos válidos

En la Tabla 28 puede observarse los valores del acuerdo de evaluación total.

Tabla 28 Acuerdo de evaluación total

Acuerdo de evaluación			
No. de inspeccionados	No. de coincidencias	Porcentaje	IC de 95%
30	22	73.33	(54.11, 87.72)

Una vez conociendo el acuerdo de concordancia, se determinó el índice Kappa de Fleiss, el cual utiliza la siguiente fórmula:

$$\kappa = \frac{p_a - p_\varepsilon}{1 - p_\varepsilon}$$

En la Tabla 29 se muestran los datos organizados para el cálculo de Kappa de Fleiss. Donde:

- **# de comparaciones:** Se refiere al número de comparaciones, en este caso son 30.
- **Suma de criterio "SÍ":** Se refiere al total de criterios "SÍ" determinadas por los evaluadores.
- **Suma de criterio "NO":** Se refiere al total de criterios "SÍ" determinadas por los evaluadores.
- **Total de criterios:** Se refiere a la suma de criterios "SÍ" y "NO".

[10] https://support.minitab.com/es-mx/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/supporting-topics/attribute-agreement-analysis/kappa-statistics-and-kendall-s-coefficients/

Tabla 29 Datos para el cálculo de Kappa de Fleiss

# de comparaciones	Suma de criterio "Sí"	Suma de criterio "No"	Total de criterios
1	5	0	5
2	0	5	5
3	5	0	5
4	5	0	5
5	5	0	5
6	5	0	5
7	1	4	5
8	5	0	5
9	4	1	5
10	1	4	5
11	5	0	5
12	1	4	5
13	5	0	5
14	4	1	5
15	5	0	5
16	5	0	5
17	5	0	5
18	0	5	5
19	4	1	5
20	5	0	5
21	4	1	5
22	1	4	5
23	5	0	5
24	0	5	5
25	5	0	5
26	5	0	5
27	5	0	5
28	0	5	5
29	5	0	5
30	5	0	5
Total general	110	40	**150**

Sustituyendo los datos en la fórmula, se determinó el Kappa de Fleiss, el cual es del **0.7272**. En la Tabla 30 se muestran la interpretación de resultados del índice Kappa de Fleiss[11].

[11]https://www.fisterra.com/mbe/investiga/kappa/kappa2.pdf

Tabla 30 Interpretación de resultados de Kappa de Fleiss

Valores del índice Kappa de Fleiss	
Valor de Kappa de Fleiss (%)	Fuerza de la concordancia
< 0.20	Pobre
0.21 – 0.40	Débil
0.40 – 0.60	Moderada
0.61 – 0.75	Buena
>0.75	Excelente

Con la interpretación de los resultados, se puede concluir que la fuerza de concordancia que se obtiene de las comparaciones entre los evaluadores en este trabajo es “Buena”.

Como ya se había mencionado, el objetivo del análisis manual y automático, es determinar el rango de los valores en los que se agrupan los porcentajes de coincidencia por cada evaluador, para esto se recurrió al intervalo de confianza, el cual muestra un rango de valores y en el cual se encuentra, con alta probabilidad, el valor real de una determinada variable. Además, éstos construyen intervalos del 95% de la probabilidad incondicional y condicional mediante la técnica Bootstrap[12].

El intervalo de confianza se calcula tomando la desviación estándar y dividiéndola por la raíz cuadrada del tamaño de la muestra, de acuerdo con la Ecuación 1:

$$\sigma x = \sigma / \sqrt{n}$$

Ecuación 1 Ecuación del intervalo de confianza

En la Figura 67 se muestra la matriz de los porcentajes arrojados por el coeficiente de similitud coseno de las 30 comparaciones.

0.622 %	**0.841 %**	**0.877 %**
0.246 %	0.530 %	0.568 %
0.611 %	0.584 %	0.559 %
0.609 %	0.741 %	0.865 %
0.693 %	0.658 %	0.724 %

[12] El bootstrapping (o bootstrap) es un método de remuestreo propuesto por Bradley Efron en 1979. Se utiliza para aproximar la distribución en el muestreo de un estadístico.

0.631 %	0.769 %	0.942 %
0.445 %	0.854 %	0.766 %
0.653 %	0.488 %	0.422 %
0.694 %	0.656 %	0.734 %
0.497 %	0.840 %	0.751 %

Figura 67 Matriz de los porcentajes arrojados del algoritmo similitud coseno

Con los valores de la matriz anterior se determinó el intervalo de confianza, quedando de la siguiente manera.

Media: 0.662
Desviación estándar: 0.155
Tamaño de la muestra: 30
Intervalo de confianza:95 %
Resultado del intervalo de confianza: **.606** (límite inferior) a .**717** (límite superior)

Del análisis anterior se puede determinar el rango de los valores en los que se agrupan los porcentajes, quedando como se muestra en la Figura 68.

Figura 68 Rango del intervalo de confianza

Como se puede observar los valores arriba del .0.717 % presentan similitud entre los capítulos, mientras que los que se encuentra por debajo de 0.606 %, no existe similitud. Cabe señalar que los valores que se encuentran entre 0.606 % y 0.717 % son tomados como "*datos impuros*" ya que se traslapan entre ellos y generan ambigüedad de decisión, por lo tanto, es necesario revisar manualmente los capítulos para determinar si existe similitud entre ellos.

Capítulo VIII

Resultados

Resultados

A continuación se muestran los resultados obtenidos del sistema desarrollado, con la finalidad de validar si el sistema está detectando correctamente los errores, por lo tanto se determinó considerar los criterios de evaluación denominados precisión y cobertura. Las ecuaciones para calcular estos valores son las siguientes[13]:

Ecuación 2:

Precisión*=TP/(TP+FP)*

Ecuación 3:

Cobertura*=TP/(TP+FN)*

Donde:

"Verdaderos Positivos" (TP): Son aquellas validaciones reportadas como correctas, y que sí lo son.

"Falsos positivos" (FP). Son aquellas validaciones reportadas como correctas, pero no lo son.

"Falsos negativos" (FN). Son aquellas validaciones reportadas como incorrectas, y que no lo son.

Cabe mencionar que cada uno de los capítulos fueron etiquetados y analizados manualmente por hablantes nativos del idioma español, posteriormente estos mismos capítulos fueron analizados por el sistema. Una vez realizados los dos análisis se compararon para calcular las medidas antes mencionadas. A continuación se presentan los resultados que se obtuvieron.

• Resultados de concordancia nominal y verbal

Para la evaluación de las reglas de concordancia nominal y verbal que se implementaron en el autómata, se analizaron 647 oraciones. En seguida se presentan los resultados de precisión y cobertura que se obtuvieron:

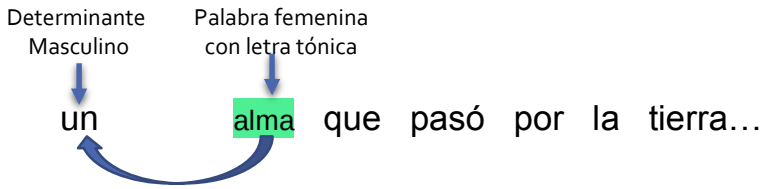
Precisión*=TP/(TP+FP)*	*515 / (515 + 92) = 0.84 %*
Cobertura*=TP/(TP+FN)*	*515 / (515 + 65) = 0.88 %*

[13] http://www.uam.es/docencia/geoteca/articulos/error/Esp/Error,%20Exactitud%20y%20Precision.htm

En el caso de los elementos detectados como falsos positivos, se debe a que algunos elementos en la oración cumplen funciones distintas a las implementadas en el autómata, por ejemplo:

El verbo de la oración “utilizando” no hace referencia al sustantivo “chistes” sino al sustantivo “detección”

Otro falso positivo que arroja el sistema, se debe a una deficiencia del etiquetado de FreeLing, el cual no reconoce al determinante que se antepone a una palabra con género femenino que empiece con una “a” o “ha” tónica y por lo tanto detecta a éste de forma incorrecta, ya que debe cambiarse a *masculino (“el”)* con este tipo de palabras, por ejemplo:

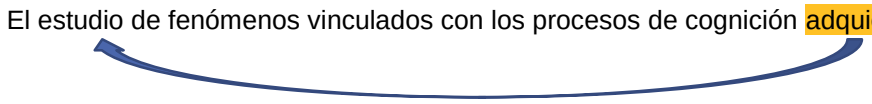


- ## Resultados de anáforas

Para la evaluación de las reglas de anáfora que se implementaron en el autómata, se analizaron 215 oraciones. En seguida se presentan los resultados de precisión y cobertura que se obtuvieron:

Precisión=*TP/(TP+FP)*	*155 / (155 + 31) = 0.83 %*
Cobertura=*TP/(TP+FN)*	*155 / (155 + 25) = 0.86 %*

En el caso de los elementos detectados como falsos positivos, se debe a que algunos elementos en la oración cumplen funciones distintas a las implementadas en el autómata, por ejemplo:



El verbo de la oración “adquiere” no hace referencia a “los procesos” sino a “el estudio”

- ## Resultados de redacción en tercera persona

Para la evaluación de la escritura en tercera persona, se analizaron 962 verbos y se obtuvieron los siguientes resultados de precisión y cobertura:

Precisión=*TP/(TP+FP)*	*912 / (912 + 37) = 0.96 %*
Cobertura=*TP/(TP+FN)*	*912 / (912 + 17) = 0.98 %*

- ## Resultados del capítulo “objetivos"

Para la evaluación de la escritura de los objetivos, se analizaron un total de 134 verbos y se obtuvieron los siguientes resultados de precisión y cobertura:

Precisión=*TP/(TP+FP)*	*120 / (120 + 5) = 0.96 %*
Cobertura=*TP/(TP+FN)*	*120 / (120 + 3) = 0.97 %*

- ## Resultados de rimas

Para la evaluación de rimas presentes en las tesis, se analizó automáticamente un total de 5920 palabras, de las cuales el sistema detecto 990 palabras como rimas en el texto. Dicho análisis fue cotejado con el análisis manual y de determinó que, aunque el sistema detecto 960 palabras con la misma terminación no todas son consideradas rimas, por lo tanto, se obtuvieron los siguientes resultados.

Precisión=*TP/(TP+FP)*	*960 / (960 +12) = 0.98 %*
Cobertura=*TP/(TP+FN)*	*960 / (960 + 9) = 0.99 %*

- ## Resultados de cacofonías

Para la evaluación de cacofonías, se analizó automáticamente un total de 4840 palabras, de las cuales 580 palabras fueron identificadas como cacofonía en el texto. Cabe señalar que este análisis fue cotejado con el análisis manual, de dicha comparación se determinó que no todas las palabras causan cacofonía.

A continuación se muestran los resultados obtenidos.

Precisión=*TP/(TP+FP)*	*540 / (540 +15) = 0.97 %*
Cobertura=*TP/(TP+FN)*	*540 / (540 +12) = 0.98 %*

- ## Resultados de repetición de vocablos

Con base en las pruebas realizadas, se midió la distancia entre las palabras repetidas con la finalidad de establecer el rango de distancia.

Para validar la opinión de los expertos y determinar que existe una correlación buena entre sus opiniones se recurrió al índice Kappa de Fleiss, en la

Tabla 32 se muestran los resultados de las comparaciones entre los hablantes nativos (evaluadores). Donde:

- *ID párrafo*: se refiere al número de identificación del párrafo que está siendo analizado.
- *ID Experto*: se refiera al número de identificación del hablante nativo del idioma español que analiza el texto.
- *Distancia*: se refiere a la distancia entre las palabras que se repiten (esta distancia se estableció en el capítulo de pruebas).
- *Índice de Kappa de Fleiss*: Se refiere al porcentaje de correlación entre las opiniones de los hablantes nativos.
- *Correlación*: Se refiere a la fuerza de coincidencia del porcentaje del índice kappa.

En la Tabla 31 Interpretación de resultados de Kappa de Fleiss se muestran la interpretación de resultados del índice Kappa de Fleiss[14].

Tabla 31 Interpretación de resultados de Kappa de Fleiss

Valores del índice Kappa de Fleiss	
Valor de Kappa de Fleiss (%)	Fuerza de la concordancia
< 0.20	Pobre
0.21 – 0.40	Débil
0.40 – 0.60	Moderada
0.62 – 0.75	Buena
>0.75	Excelente

[14] https://www.fisterra.com/mbe/investiga/kappa/kappa2.pdf

Tabla 32 Distancia entre las palabras repetidas

Posición de distancia entre palabras repetidas							
ID párrafo	Distancia de palabras repetidas				Índice de Kappa de Fleiss	Correlación	Rango
	ID experto						
	ID 1	ID 2	ID 3	ID 4			
1	3 7 14	3 7 14	3 7 14	3 7 0	88 %	Excelente	3 hasta 14
2	2 5	2 5	2 5	2 0	80 %	Excelente	2 hasta 5
3	3 4 7 9	3 4 7 0	3 4 0 0	3 4 7 9	76 %	Excelente	3 hasta 9
4	2 13	2 0	2 13	2 13	80 %	Excelente	2 hasta 13
5	2 6 17	2 6 0	2 6 17	2 6 17	88 %	Excelente	2 hasta 17
6	13	0	13	13	55 %	Regular	0 hasta 13
7	3 11 13	3 11 0	0 0 0	3 11 13	50 %	Regular	3 hasta 13
8	2 4 19	2 4 0	2 4 0	2 4 19	76 %	Excelente	2 hasta 19
9	4 5	0 0	4 0	4 5	40 %	Regular	4 hasta 5
10	2 3	2 0	0 0	2 3	40 %	Regular	2 hasta 3
				Total	**67 %**	**Buena**	**2 hasta 19**

Nota: Los cálculos del índice Kappa de Fleiss de cada párrafo se encuentra en el Anexo 1.

Con la interpretación de los resultados, se puede concluir que la fuerza de concordancia que se obtiene de las comparaciones entre los hablantes nativos (evaluadores) es “Buena”. Por lo tanto, las pruebas son confiables y se puede concluir que la distancia entre las palabras que se repiten es de **2 a 19,** es decir la

distancia mínima en la que se encuentra una palabra repetida es de 2 y la máxima de 19.

- ## Resultados del módulo "Identificación de similitud semántica entre capítulos"

Para los resultados de esta sección se analizaron 10 tesis automáticamente por el algoritmo de similitud coseno, posteriormente los porcentajes arrojados por el algoritmo se compararon con los rangos de valores que se establecieron en las pruebas del capítulo 3.2, de esta forma se evaluó la eficacia del algoritmo.

A continuación, en las siguientes tablas se muestran las 10 comparaciones manuales y automáticas que se realizaron, se puede observar del lado derecho los porcentajes arrojados por el algoritmo y del lado izquierdo el análisis realizado por los hablantes nativos los cuales marcan con un "*Sí*" si existe similitud entre los capítulos y un "*NO*" si no existe. En dichas tablas se pueden observar los siguientes parámetros:

- **ID de comparación:** Se refiere al número de identificación de la comparación entre capítulos que está siendo analizada.
- **Similitud encontrada por hablantes:** Este campo está dividido en:
  - **ID hablantes:** Se refiera al número de identificación del hablante nativo del idioma español que analiza el texto, los cuales marcan con un "Sí" si existe similitud entre los capítulos y un "NO" si no existe.
- **Mayoría:** Este campo se refiere a la mayoría de "SÍ" o "NO" que se encuentran en la columna "Similitud encontrada por hablantes". Cabe señalar que si existe un empate de "SÍ" o "NO" se coloca un guion (-).
- **Porcentaje de similitud del algoritmo:** se refiere al valor que arroja el algoritmo de similitud coseno.
- **Resultado:** Se determina un "Sí" o un "No" con base al rango establecido en la fase de pruebas del capítulo 3.2.
-

Igualmente, en cada una de las tablas se aprecian tres colores, blanco, verde y anaranjado, los cuales corresponden a:

Comparaciones correctas: Se refiere a la compatibilidad entre la decisión de los evaluadores (Mayoría) y la decisión del algoritmo en función del rango determinado (Resultado) que debe ser superior a 0.717 %.

Comparaciones erróneas: Se refiere a la incompatibilidad entre la decisión de los evaluadores (Mayoría) y la decisión del algoritmo en función del rango determinado (Resultado) que debe ser menor a 0.606 %.

Comparaciones ambiguas: No se puede determinar si es una comparación correcta o incorrecta, debido a que se encuentra dentro del rango de los datos impuros (0.605 % 0.716 %)

Resultados de la comparación entre “Título - Objetivo”

En la Tabla 33 se muestra la comparación de los capítulos “Título – Objetivo”, se compara el análisis manual con el automático, se determina si existe similitud a partir del rango establecido en al capítulo 3.2.

Tabla 33 Comparación “Título – Objetivo”

Comparación “Título - Objetivo”						
ID de comparación	Similitud encontrada por hablantes			Mayoría	Porcentaje de similitud del algoritmo	Resultado
	ID hablantes					
	ID 1	ID 2	ID 3			
1	SÍ	SÍ	SÍ	SÍ	0.801 %	SÍ
2	NO	NO	NO	NO	0.350 %	NO
3	NO	NO	NO	NO	0.289 %	NO
4	SÍ	NO	NO	NO	0.885 %	SÍ
5	NO	SÍ	SÍ	SÍ	0.604 %	NO
6	NO	NO	NO	NO	0.549 %	NO
7	NO	NO	SÍ	NO	0.677 %	NO
8	SÍ	NO	SÍ	SÍ	0.595 %	NO
9	SÍ	NO	NO	NO	0.530 %	NO
10	NO	NO	NO	NO	0.775 %	SÍ

Resultados de la comparación entre “Planteamiento del problema – Objetivo”

En la Tabla 34 se muestra la comparación de los capítulos “Planteamiento del problema – Objetivo”, se compara el análisis manual con el automático, se determina si existe similitud a partir del rango establecido en al capítulo 3.2.

Tabla 34 Comparación “Planteamiento del problema – Objetivo”

Comparación “Planteamiento del problema - Objetivo”						
ID de comparación	Similitud encontrada por hablantes			Mayoría	Porcentaje de similitud del algoritmo	Resultado
	ID hablantes					
	ID 1	ID 2	ID 3			
1	SÍ	NO	NO	NO	0.650 %	NO
2	SÍ	SÍ	SÍ	SÍ	0.840 %	SÍ
3	NO	NO	NO	NO	0.523 %	NO
4	SÍ	SÍ	SÍ	SÍ	0.881 %	SÍ
5	NO	NO	SÍ	NO	0.881 %	SÍ
6	SÍ	SÍ	SÍ	SÍ	0.735 %	SÍ
7	NO	SÍ	SÍ	SÍ	0.803 %	SÍ
8	SÍ	SÍ	SÍ	SÍ	0.778 %	SÍ
9	SÍ	SÍ	SÍ	SÍ	0.705 %	- (NO)
10	SÍ	SÍ	SÍ	SÍ	0.812 %	SÍ

*Debido a que el porcentaje se encuentra dentro de los valores conocidos como “datos impuros”, se toma el valor como negativo para los cálculos de cobertura y precisión.

Resultados de la comparación entre “Objetivo - Conclusión”

En la Tabla 35 se muestra la comparación de los capítulos “Objetivo - Conclusión”, se compara el análisis manual con el automático, se determina si existe similitud a partir del rango establecido en al capítulo 3.2.

Tabla 35 Comparación “Objetivo - Conclusión”

Comparación “Objetivo -Conclusión”						
ID de comparación	Similitud encontrada por hablantes			Mayoría	Porcentaje de similitud del algoritmo	Resultado
	ID hablantes					
	ID 1	ID 2	ID 3			
1	SÍ	SÍ	SÍ	SÍ	0.877 %	SÍ
2	SÍ	SÍ	SÍ	SÍ	0.841 %	SÍ
3	NO	NO	NO	NO	0.326%	NO
* 4	SÍ	NO	SÍ	SÍ	0.608 %	- (NO)
5	NO	NO	NO	NO	0.349 %	NO
6	NO	NO	NO	NO	0.463 %	NO
7	SÍ	NO	SÍ	SÍ	0.505 %	NO
8	SÍ	SÍ	SÍ	SÍ	0.925 %	SÍ
9	SÍ	SÍ	SÍ	SÍ	0.836 %	SÍ
10	SÍ	NO	NO	NO	0.778 %	SÍ

*Debido a que el porcentaje se encuentra dentro de los valores conocidos como “datos impuros”, se toma el valor como negativo para los cálculos de cobertura y precisión. Sí se desea saber si existe similitud entre los capítulos se debe examinar manualmente.

A continuación se muestran los resultados de cobertura y precisión:

Verdaderos Positivos (TP): **23**

Falsos positivos (FP): 4

Falsos negativos (FN): 3

Por lo tanto, los valores de cobertura y precisión quedaron de la siguiente manera:

Precisión=*TP/(TP+FP)* ***Precisión***=*23/(23+4) = 85 %*

Cobertura=*TP/(TP+FN)* ***Cobertura***=*23/(23+3) = 88 %*

- ## Promedio de precisión y cobertura del sistema

En la Tabla 36 se muestran los fenómenos lingüísticos que se analizan y los resultados de cobertura y precisión de cada uno de ellos.

Tabla 36 Precisión y cobertura de los diferentes fenómenos lingüísticos

Fenómeno lingüístico	Precisión	Cobertura
Concordancia nominal y verbal	84 %	88 %
Anáfora	83 %	86 %
Redacción en tercera persona	96 %	98 %
Redacción del capítulo objetivos	96 %	97 %
Rimas	98 %	99 %
Cacofonías	97 %	98 %
Similitud entre capítulos	85 %	88 %

Por lo tanto, el promedio de cobertura y precisión es el siguiente:

Precisión = 91 % Cobertura = 93 %

Capítulo IX

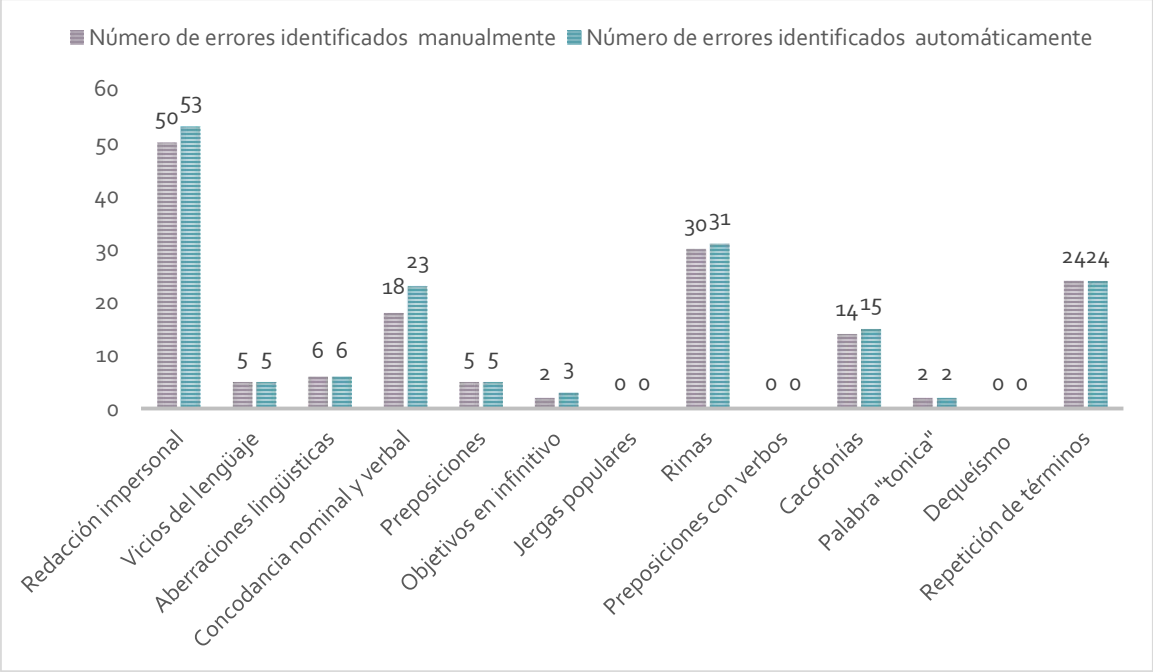
Conclusión

Conclusiones

El sistema desarrollado permite identificar errores gramaticales y de estilo, lo cual facilitará la revisión de las tesis tanto a estudiantes como a revisores, disminuyendo el tiempo de detección de errores y proporcionando sugerencias de corrección para mejorar la calidad del escrito.

Como se ha mencionado en el transcurso de este trabajo, se realizaron dos tipos de análisis (manual y automático). Posteriormente, dichos análisis fueron cotejados con la finalidad de conocer si los resultados de los errores identificados coincidían, esto puede observarse en la Gráfica 1.

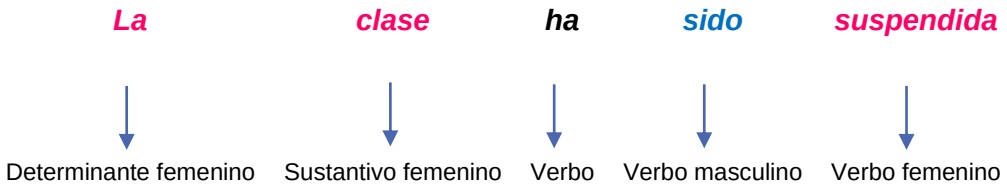
Gráfica 1 Número de errores gramaticales y de estilo identificados



Como se observa en la Gráfica 1, los errores que tienen que ver con el uso de jergas populares, preposiciones con verbos y dequeísmo son de frecuencia nula en las tesis, por el contrario, errores de redacción impersonal, rimas y repetición de términos son más frecuentes en la escritura de la tesis.

Cabe destacar que el sistema identifica los siguientes tipos de error:

- Error de concordancia con algunos verbos como “sido”, “mantenido” entre otros, a pesar de ser verbos de género masculino son utilizados en expresiones que tienen sustantivos y determinantes femeninos, por ejemplo:



- La palabra “ser”, genera ambigüedad y su significado depende del contexto, por ejemplo:

  En la oración *“El **ser** humano…”* la palabra *“ser”* cumple la función de sustantivo, pero en la oración *“el problema puede **ser** resuelto…”* cumple la función de verbo.

- Regla con determinante masculino + sustantivo femenino que empiezan con “a” o “ha” tónica + adjetivo femenino.

El águila blanca

También se observó durante la realización este trabajo, que existen verbos que se utilizan con más de una preposición, dependiendo el contexto en el que son empleados, por lo tanto, no son identificados por el sistema; a continuación se muestran ejemplos de este fenómeno lingüístico.

- Verbos que deben ir acompañados de ciertas preposiciones, se debe analizar el contexto, para determinar que preposición debe emplearse; algunos ejemplos son los siguientes:

“no son capaces de ***pensar de*** manera abstracta…”

“Debemos ***pensar en*** los usuarios…”

Como se puede observar en las oraciones, el verbo “*pensar*” se utiliza con las preposiciones “*de*” y “*en*” dependiendo el contexto. Lo mismo sucede con el siguiente ejemplo:

*"Juan se **casó con** Carla"*

*"Juan se **casó en** Cuernavaca"*

En la primera oración la preposición "*con*" hace referencia a la persona con la que se casó, mientras que en la segunda, la preposición "*en*" hace referencia al lugar".

En estos dos ejemplos se demuestra, como ya se había mencionado, que existen verbos que pueden ir acompañados con más de una preposición pero para determinar cuál es la correcta se debe analizar el contexto de las oraciones, desafortunadamente los alcances del sistema no permiten determinar el contexto, sin embargo; se tiene en consideración para trabajo futuro.

También se observó que algunos criterios de estilo como las rimas y cacofonías poseen diferentes características que las que se conocen popularmente, como la distancia, el número de palabras y letras que generan errores rimas y cacofonías.

Otro factor que altera los resultados del análisis, tiene que ver con una mala ortografía, por ejemplo, las palabras "esta", "ésta" y "está" no significan lo mismo, ya que, "esta" se emplea como adjetivo "esta metodología fue mal diseñada", "ésta" se emplea como pronombre "ésta fue empleada" y "está" se emplea como verbo "la metodología está mal diseñada". Es por ello que al estar mal acentuada alguna palabra o escribirla incorrectamente se alteran los resultados ya que la función que cumplen en la oración no es la adecuada.

Por otro lado, también se generaron diversos recursos que permitieron alimentar al sistema para identificar errores gramaticales y de estilo, estos recursos son listas de: preposiciones, preposiciones con verbos, palabras tónicas, vulgarismos, extranjerismos, barbarismos, pleonasmos, vocabulario rebuscado, jergas populares y modismos (Anexo 1).

Recursos generados

Durante el desarrollo de este trabajo se generaron diferentes recursos, los cuales se encuentran en formato plano, por lo tanto, la manipulación de los datos es sencilla. A continuación se enlistan los recursos generados:

- Recurso de preposiciones

  Este recurso cuenta con 38 preposiciones divididas en dos listas, en la lista 1 se encuentran las preposiciones de forma incorrecta y en la lista 2 las preposiciones correctas.

- Recurso de preposiciones con verbos

  Este recurso cuenta con 159 preposiciones divididas en cuatro listas, la lista 1 contiene las preposiciones que deben ir acompañadas de la preposición "***a"***, en la lista 2 los de la preposición *"de",* la tercera los de la preposición "*con"* y por último los que deben ir con "*en"*.

- Recurso de palabras tónicas

  Este recurso contiene 50 palabras que inicia con "*a"* y "*ha"* tónica, por ejemplo: *águila, agua, área, hambre*, etc.

- Recurso de vicios del lenguaje

  El recurso de vicios del lenguaje está dividido en 4 categorías, la primera categoría es de los barbarismos, la segunda de pleonasmos, la tercera de vulgarismos y la cuarta de extranjerismos. El recurso está formado por un total de 180 palabras y cada categoría está dividido en dos listas, la primera lista contiene los errores de vicios del lenguaje y la segunda contiene la forma de evitarlo.

- Recurso de Jergas populares

  Este recurso está formado de palabras que son consideradas jergas populares como "*cacho", "meramente", entre otros*. El número total de jergas populares que conforman este recurso es de 418 palabras.

- Recurso de Aberraciones lingüísticas

  El recurso de aberraciones lingüísticas, está dividido en dos categorías, la primera es verbosidad y la segunda, vocabulario rebuscado. El total de

palabras que conforman el recurso es de 36 palabras, y cada categoría esta divida en dos listas, la primera lista contiene las palabras que generan errores y la segunda la forma correcta de escribir.

Cabe señalar que cada uno de los recursos puede ampliarse de manera sencilla debido al formato con el que están creados.

Anexo I

Cálculo de calidad en un texto

El cálculo de la calidad se realizó a través de dos criterios, los cuales serán explicados a continuación:

- El primer criterio de calidad, se deriva del tamaño de los textos:

Se tomó en cuenta la cantidad de palabras correctas y erróneas que conforman el texto y se utilizó la siguiente formula [11]:

NPE/NTP=IE %

(Número de palabras erróneas/ número total de palabras) / 100 = indicador de error

- Segundo criterio, coherencia, consistencia y legibilidad:

Se toma en cuenta los siguientes factores:

- Presencia de reglas gramaticales en las oraciones
- Presencia de redundancias en las oraciones
- Presencia de cohesión, es decir, que las oraciones estén bien unidas gramaticalmente [30].

Así mismo se desarrollaron 14 rúbricas para evaluar la calidad de la escritura, las cuales describen una variedad de características deseables en los textos argumentativos. En la Tabla 37 se puede apreciar dichas rúbrica que se aplicaron a 20 capítulos de “Introducción” de tesis, ya que este capítulo es uno de los que poseen mayor extensión y existe mayor probabilidad de encontrar errores.

Tabla 37 Rúbricas de evaluación

Id rúbrica	Rúbrica	Id capítulos de tesis Valor= 0=No 1= Sí																			
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	Mal uso de preposiciones	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
2	Mal uso de preposiciones con verbo	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	Mal uso de dequeísmo	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0
4	Mal uso de Concordancia nominal	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	Mal uso Concordancia verbal	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1
6	Mal uso de palabras tónicas	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
7	Uso de vicios del lenguaje	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	1
8	Redacción de capítulos en primera persona	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1
9	Redacción de objetivos que no sean en tiempo infinitivo	1	1	0	0	0	1	0	1	1	0	1	1	1	0	1	1	1	0	1	0
10	Uso de rimas	1	0	1	0	0	1	0	0	1	0	1	1	0	0	1	1	0	1	0	1
11	Uso de cacofonías	1	1	1	1	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1	0
12	Uso de jergas populares	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
13	Uso de aberraciones lingüísticas	1	0	1	0	0	1	1	0	0	0	1	0	1	1	1	0	0	0	1	1
14	Repetición de vocablos	1	1	0	1	0	0	1	1	0	0	0	1	1	0	0	0	1	1	1	0

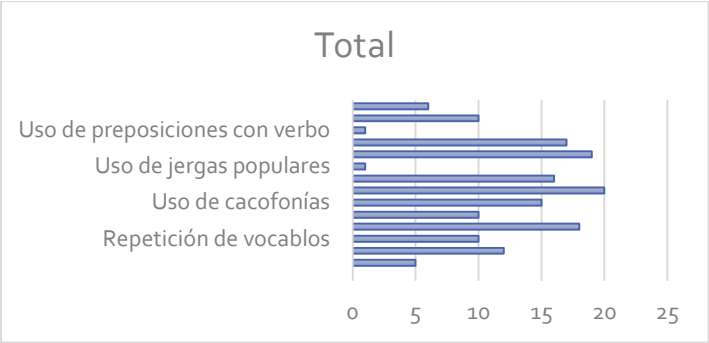
En la tabla anterior se muestran las rubricas con el valor 0 para No y 1 para Sí. Cabe señalar que las métricas están fundamentadas por el Instituto Nacional para la Evaluación de la Educación (INNE)[15], con la finalidad de lograr altos grados de confiabilidad en la evaluación de la escritura. La evaluación de las rubricas se realizó de forma automática mediante el sistema desarrollado en este trabajo de tesis.

Se utilizó la Teoría de la Medida16 para evaluar la calificación global en la prueba de escritura, para ellos se hizo la sumatoria de las 14 rúbricas, cuya mayor puntuación es 11 puntos. Así mismo, esta teoría sugiere que se calcule de dos a tres parámetros básicos de las pruebas para una confiabilidad mayor. Dado que las 14 rúbricas se clasificaron de manera distinta (0 y 1 puntos), se consideró como un "Sí" las respuestas con valor igual a uno y con un "No" las que tienen valor a cero.

[15] Medidas de evaluación de la escritura, realizadas por La Dra. Margarita Peón Zapata y la Mtra. Sara Rivera ambas de la Dirección de Pruebas y Medición del INNE.

[16] Esta teoría pretende explicar la manera en que a partir de un valor de test medido de una persona se puede concluir el «valor verdadero» de la manifestación de la característica que se quiere medir.

Como se puede observar en la Gráfica 2 siguiente la rúbrica con mayor puntuación es concordancia nominal y la de menor puntuación es jergas populares.



Gráfica 2 Puntuaciones de las rúbricas

Cálculo de Alfa de Cronbach

La confiabilidad de las 14 rubricas se llevó a cabo mediante el Coeficiente de Alfa de Cronbach el cual permite estimar la fiabilidad de un instrumento de medida a través de un conjunto de ítems que se espera que midan el mismo constructo o dimensión teórica [31]. En la Tabla 38 se observa el cálculo del Coeficiente de Alfa de Cronbach de las 14 rúbricas.

Tabla 38 Cálculo del Coeficiente de Alfa de Cronbach

Id rúbrica	Rúbrica	Id capítulos de tesis Valor= 0=No 1= Sí																				Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	Mal uso de preposiciones	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	17
2	Mal uso de preposiciones con verbo	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
3	Mal uso de dequeísmo	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0	17
4	Mal uso de Concordancia nominal	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
5	Mal uso Concordancia verbal	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	18
6	Mal uso de palabras tónicas	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	19
7	Uso de vicios del lenguaje	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	1	6
8	Redacción de capítulos en primera persona	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	5
9	Redacción de objetivos que no sean en tiempo infinitivo	1	1	0	0	0	1	0	1	1	0	1	1	1	0	1	1	1	0	1	0	12
10	Uso de rimas	1	0	1	0	0	1	0	0	1	0	1	1	0	0	1	1	0	1	0	1	10
11	Uso de cacofonías	1	1	1	1	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1	0	15
12	Uso de jergas populares	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
13	Uso de aberraciones lingüísticas	1	0	1	0	0	1	1	0	0	0	1	0	1	1	1	0	0	0	1	1	10
14	Repetición de vocablos	1	1	0	1	0	0	1	1	0	0	0	1	1	0	0	0	1	1	1	0	10
Varianza		0.18	0.26	0.26	0.26	0.26	0.25	0.26	0.26	0.25	0.27	0.25	0.26	0.25	0.25	0.25	0.26	0.27	0.25	0.25	0.27	41.81

El cálculo de Alfa de Cronbach se realiza a través de la siguiente ecuación:

$$\propto = \frac{K}{K-1}\left[1 - \frac{\sum Vi}{Vt}\right]$$

Sección 1 Sección 2

Ecuación 4 Alfa de Cronbach

Donde K es el número total de ítems (rúbrica), ∑Vi es la suma de la varianza de los ítems y Vt es la varianza total.

Por lo tanto:

k	20
suma Varianza	5.082
Varianza total	41.80
sección 1	1.053
sección 2	0.878
absoluto s2	0.878
α	0.925

Los rangos del coeficiente de Alfa de Cronbach, oscilan entre los siguientes valores (ver Tabla 39).

Tabla 39 Rangos del coeficiente de Alfa de Cronbach

Valor	Confiabilidad
<0.60	Mala
0.61-0.79	Regular
>0.80	Buena

Como puede apreciar el coeficiente de Alfa de Cronbach es de 0.92 % por lo tanto tiene una buena confiabilidad.

Cálculo de calidad de la escritura

Una vez determinado que las pruebas de calidad de la escritura, se realizó la indización de la calidad de las palabras de cada capítulo de tesis, por lo tanto se tomó en cuenta la siguiente fórmula[17]:

$$IC\ (\%) = (100*A)/A+M$$

Donde A son las palabras que presentaban errores y M es el total de palabras extraídas de los 20 capítulos de las tesis analizadas.

Se realizaron tres evaluaciones (ver Tabla 40):

1.- Capítulo con mayor cantidad de 1 (Id Capítulo 1)

2.- Capítulo con menos cantidad de 1 (Id Capítulo 14)

3.- Capítulo con casi la misma cantidad de 1 y 0 (Id Capítulo 19)

Tabla 40 Evaluaciones de capítulos

Id rúbrica	Rúbrica	Id capítulos de tesis Valor= 0=No 1= Sí																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Mal uso de preposiciones	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
2	Mal uso de preposiciones con verbo	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	Mal uso de dequeísmo	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0
4	Mal uso de Concordancia nominal	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	Mal uso Concordancia verbal	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1
6	Mal uso de palabras tónicas	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
7	Uso de vicios del lenguaje	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	1
8	Redacción de capítulos en primera persona	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1
9	Redacción de objetivos que no sean en tiempo infinitivo	1	1	0	0	0	1	0	1	1	0	1	1	1	0	1	1	1	0	1	0
10	Uso de rimas	1	0	1	0	0	1	0	0	1	0	1	1	0	0	1	1	0	1	0	1
11	Uso de cacofonías	1	1	1	1	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1	0
12	Uso de jergas populares	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
13	Uso de aberraciones lingüísticas	1	0	1	0	0	1	1	0	0	0	1	0	1	1	1	0	0	0	1	1
14	Repetición de vocablos	1	1	0	1	0	0	1	1	0	0	0	1	1	0	0	0	1	1	1	0
Total		11	8	8	7	6	9	8	8	9	7	9	8	9	5	9	8	7	9	9	7

- [17] ABAD GARCÍA, F. Investigación evaluativa en documentación. Valencia : Universitat de Valencia, 1997. p. 130-132

Se calcularon el total de palabras correctas e incorrectas de los capítulos seleccionados (Tabla 41).

Tabla 41 Calculo de palabras correctas e incorrectas

Id capítulo	Palabras correctas	Palabras erróneas (A)	Palabras totales (M)
1	3510	14	3524
14	1887	14	1901
19	2042	25	2067
Total	34498	532	35030

Sustituyendo los valores de la tabla en la fórmula para evaluar la calidad de las palabras de cada capítulo de tesis, se obtienen los siguientes resultados:

$$IC\ (\%) = (100*A)/A+M$$

Id capítulo	A	M	IC (%)
14	14	1901	0.731
19	32	592	5.128
1	14	3524	0.396

Tomando en cuenta los parámetros de medición que se aprecian en la Tabla 42, se determina los siguientes resultados [2]:

Tabla 42 Parámetros de medición de calidad en la escritura

Valor	Confiabilidad
<0.40	Mala
0.41-0.59	Regular
>0.60	Buena

Por lo tanto, según las estadísticas, la calidad de los escritos analizados para la evaluación del id capítulo 1 es de mala, para el id capítulo 14 es buena y para el id capítulo 19 es regular.

Patrones de concordancia nominal y verbal

A continuación en la Tabla 43 se muestran los patrones de concordancia nominal y verbal:

Tabla 43 Patrones de concordancia nominal y verbal

Nota: las etiquetas de los patrones están compuestas de la siguiente información:

#	Patrones	#	Patrones
1.	DFS + NFS	**24.**	DMP + NMP + AMP + VP
2.	DMS + NMS	**25.**	DFS + NFS + VS + AFS
3.	DFP + NFP	**26.**	DMS + NMS + VS + AMS
4.	DMP + NMP	**27.**	DFP + NFP + VP + AFP
5.	DFS + NFS + VS	**28.**	DMP + NMP + VP + AMP
6.	DMS + NMS + VS	**29.**	DFS + VSF + NFS
7.	DFP + NFP + VP	**30.**	DMS + VSM + NMS
8.	DMP + NMP + VP	**31.**	DFP + VPF + NFP
9.	DFS + NFS + AFS	**32.**	DMP + VPM + NMP
10.	DMS + NMS + AMS	**33.**	DFS + NFS + VSF
11.	DFP + NFP + AFP	**34.**	DMS + NMS + VSM
12.	DMP + NMP + AMP	**35.**	DFP + NFP + VPF
13.	DFS + AFS + NFS	**36.**	DMP + NMP + VPM
14.	DMS + AMS + NMS	**37.**	DFS + PRCS + VS
15.	DFP + AFP + NFP	**38.**	DMS + PRCS + VS
16.	DMP + AMP + NMP	**39.**	DFP + PRCP + VP
17.	DFS + AFS + NFS + VS	**40.**	DMP + PRCP + VP
18.	DMS + AMS + NMS + VS	**41.**	DFS + PRCS + VS + AFS
19.	DFP + AFP + NFP + VP	**42.**	DMS + PRCS + VS + AMS
20.	DMP + AMP + NMP + VP	**43.**	DFP + PRCP + VP + AFP
21.	DFS + NFS + AFS + VS	**44.**	DMP + PRCP + VP + AMP
22.	DMS + NMS + AMS + VS	**45.**	DMS + NMS + VAS + VSM + VSM
23.	DFP + NFP + AFP + VP	**46.**	DFS + NFS + VAS + VSM + VSF

Etiqueta D	Explicación Determinante

A	Adjetivo
N	Sustantivo
V	Verbo
F	Femenino
M	Masculino
S	Singular
P	Plural
PR	Pronombre
VAS	Verbo Auxiliar Singular

Kappa de Fleiss del módulo "Repetición de vocablos"

A continuación se presentan los resultados del índice de Kappa de Fleiss de cada uno de los párrafos que analizaron los hablantes nativos. Cabe recordar que la finalidad de este análisis es medir la concordancia de los evaluadores, mientras más alto el porcentaje de concordancia exista significa que hay una correlación de opiniones alta.

El cálculo del índice Kappa de Fleiss se realizó con la ayuda de Minitab[18], el cual es un programa de computadora diseñado para ejecutar funciones estadísticas básicas y avanzadas.

A continuación se muestran los cálculos de Kappa de Fleiss de cada uno de los párrafos.

Párrafo 1:

Posición de distancia entre palabras repetidas				
ID párrafo	ID experto			
	ID 1	ID 2	ID 3	ID 4
	3	3	3	3
1	7	7	7	7
	14	14	14	

[18] http://www.minitab.com/es-mx/

Estadísticos Kappa de Fleiss

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
3	1.00000	0.288675	3.46410	0.0003
0	*	*	*	*
14	0.70000	0.288675	2.42487	0.0077
7	1.00000	0.288675	3.46410	0.0003
General	0.88462	0.196900	4.49271	0.0000

Párrafo 2:

Posición de distancia entre palabras repetidas				
ID párrafo	ID experto			
	ID 1	ID 2	ID 3	ID 4
2	2	2	2	2
	5	5	5	0

Estadísticos Kappa de Fleiss

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
0	*	*	*	*
2	1.00000	0.353553	2.82843	0.0023
5	0.66667	0.353553	1.88562	0.0297
General	0.80000	0.331662	2.41209	0.0079

Párrafo 3:

Posición de distancia entre palabras repetidas				
ID párrafo	ID experto			
	ID 1	ID 2	ID 3	ID 4
3	3	3	3	3
	4	4	4	4
	7	7	0	7
	9	0	0	9

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
0	*	*	*	*
3	1.00000	0.250000	4.00000	0.0000
4	1.00000	0.250000	4.00000	0.0000
7	0.71429	0.250000	2.85714	0.0021
9	0.42857	0.250000	1.71429	0.0432
General	0.76000	0.137356	5.53307	0.0000

Párrafo 4:

Posición de distancia entre palabras repetidas				
ID párrafo	ID experto			
	ID 1	ID 2	ID 3	ID 4
4	2	2	2	2
	13	0	13	13

Estadísticos Kappa de Fleiss

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
0	*	*	*	*
13	0.66667	0.353553	1.88562	0.0297
2	1.00000	0.353553	2.82843	0.0023
General	0.80000	0.331662	2.41209	0.0079

Párrafo 5:

Posición de distancia entre palabras repetidas				
ID párrafo	ID experto			
	ID 1	ID 2	ID 3	ID 4
5	2	2	2	2
	6	6	6	6
	17		17	17

Estadísticos Kappa de Fleiss

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
2	1.00000	0.288675	3.46410	0.0003
0	*	*	*	*
17	0.70000	0.288675	2.42487	0.0077
6	1.00000	0.288675	3.46410	0.0003
General	0.88462	0.196900	4.49271	0.0000

Párrafo 6:

Posición de distancia entre palabras repetidas				
ID párrafo	ID experto			
	ID 1	ID 2	ID 3	ID 4
6	13	0	13	13

Estadísticos Kappa de Fleiss

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
0	0.555556	0.235702	2.35702	0.0092
13	0.555556	0.235702	2.35702	0.0092

Párrafo 7:

ID párrafo	Posición de distancia entre palabras repetidas			
	ID experto			
	ID 1	ID 2	ID 3	ID 4
7	3	3	0	3
	11	11	0	11
	13	0	0	13

Estadísticos Kappa de Fleiss

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
3	0.700000	0.288675	2.42487	0.0077
0	*	*	*	*
11	0.700000	0.288675	2.42487	0.0077
13	0.400000	0.288675	1.38564	0.0829
General	0.509615	0.190172	2.67975	0.0037

Párrafo 8:

ID párrafo	Posición de distancia entre palabras repetidas			
	ID experto			
	ID 1	ID 2	ID 3	ID 4
8	2	2	2	2
	4	4	4	4
	19	0	0	19

Estadísticos Kappa de Fleiss

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
2	1.00000	0.288675	3.46410	0.0003
4	1.00000	0.288675	3.46410	0.0003
0	*	*	*	*
19	0.40000	0.288675	1.38564	0.0829
General	0.76923	0.189401	4.06138	0.0000

Párrafo 9:

Posición de distancia entre palabras repetidas				
ID párrafo	ID experto			
	ID 1	ID 2	ID 3	ID 4
9	4	0	4	4
	5	0	0	5

Estadísticos Kappa de Fleiss

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
0	*	*	*	*
4	0.666667	0.353553	1.88562	0.0297
5	0.333333	0.353553	0.94281	0.1729
General	0.400000	0.308221	1.29777	0.0972

Párrafo 10:

Posición de distancia entre palabras repetidas				
ID párrafo	ID experto			
	ID 1	ID 2	ID 3	ID 4
10	2	2	0	2
	3	0	0	3

Estadísticos Kappa de Fleiss

Respuesta	Kappa	Error estándar de Kappa	Z	P(vs > 0)
0	*	*	*	*
2	0.666667	0.353553	1.88562	0.0297
3	0.333333	0.353553	0.94281	0.1729
General	0.400000	0.308221	1.29777	0.0972

Errores detectados por la aplicación web

A continuación, se muestran los diferentes tipos de errores detectados por la aplicación Web.

Errores de Preposiciones

En las Figura 69 y Figura 70 se muestran la vista de los errores de preposiciones que se detectan en la interfaz Web.

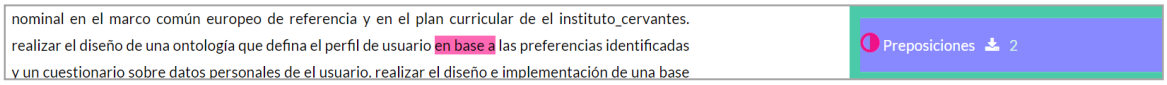


Figura 69 Errores de preposiciones analizados por la aplicación Web.

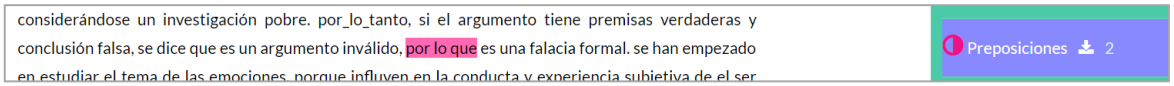


Figura 70 Errores de preposiciones analizados por la aplicación Web.

Errores de Preposiciones con verbos

En la Figura 71, se muestra el resultado de los errores que se detectan en la interfaz Web.

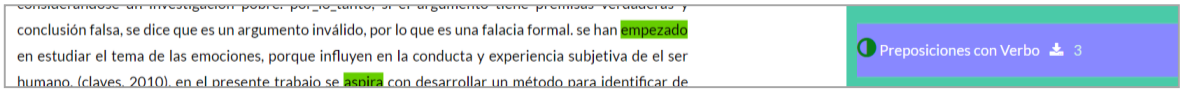


Figura 71 Errores de preposiciones con verbos analizados por la aplicación Web.

Errores de Dequeísmo

En la Figura 72, se observa la detección de errores en la plataforma Web, con respecto a los errores de dequeísmo.

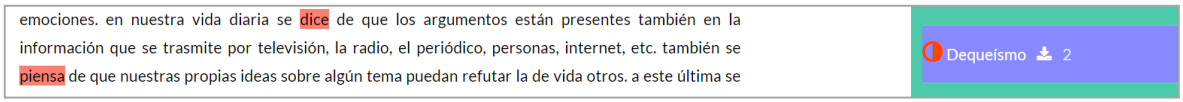


Figura 72 Errores de dequeísmo analizados por la aplicación Web.

Errores de Concordancia nominal y verbal

En la Figura 73, se observa la detección de errores de concordancia nominal y verbal.

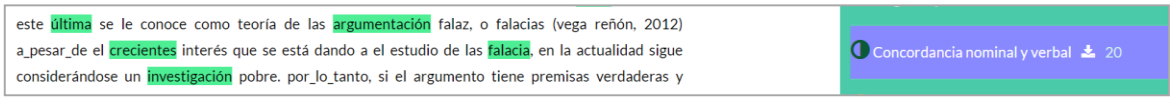


Figura 73 Errores de concordancia nominal y verbal analizados por la aplicación Web.

Errores de anáfora

En la Figura 74, se observa la detección de errores de anáfora.

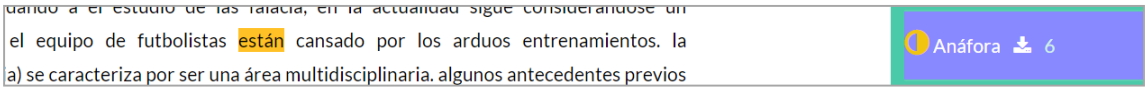


Figura 74 Errores de anáforas analizados por la aplicación Web.

Errores de palabras tónicas

En la Figura 75, se observa la detección de errores en la plataforma Web.

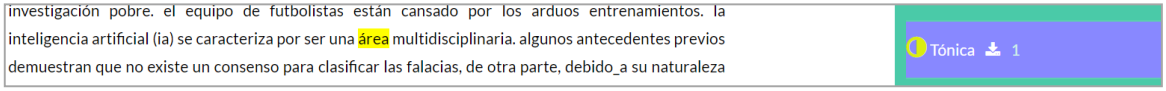


Figura 75 Errores de palabras tónicas analizados por la aplicación Web.

Errores de vicios del lenguaje

En la Figura 76, se muestra el tipo de errores llamados “vicios del lenguaje” que se detectan en la interfaz Web.

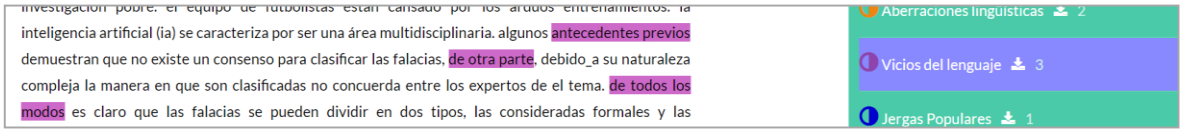


Figura 76 Errores de vicios del lenguaje analizados por la aplicación Web.

Errores de redacción impersonal

En la Figura 77, se muestra el resultado que se detectan en la interfaz Web.

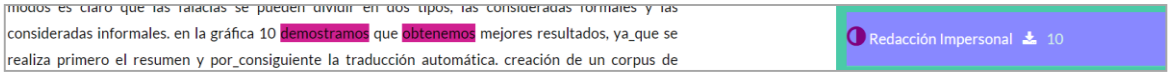


Figura 77 Errores de redacción impersonal analizados por la aplicación Web.

Errores de rimas

En la Figura 78, se observa la detección de errores en la plataforma Web, con respecto a los errores de rimas.

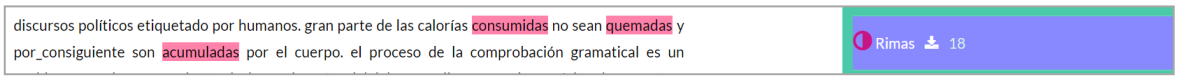


Figura 78 Errores de rimas analizados por la aplicación Web.

Errores de cacofonía

En la Figura 79, se observa la detección de errores de cacofonías en la plataforma Web.

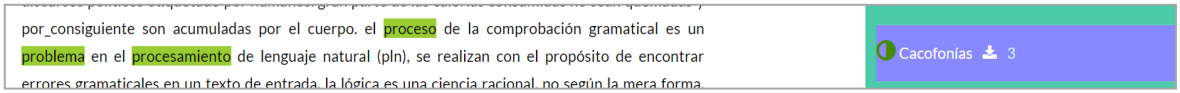


Figura 79 Errores de cacofonía analizados por la aplicación Web.

Errores de jergas populares

En la Figura 80, se muestra el resultado de los errores que se detectan en la interfaz Web.

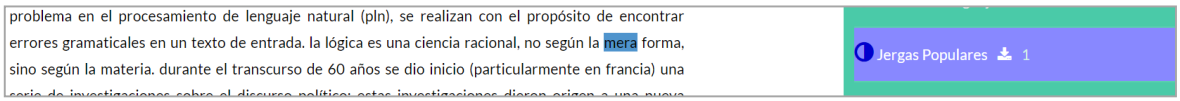


Figura 80 Errores de jergas populares analizados por la aplicación Web.

Errores de reiteración de vocablo

En la Figura 81, se muestra el resultado de los errores que se detectan en la interfaz Web.

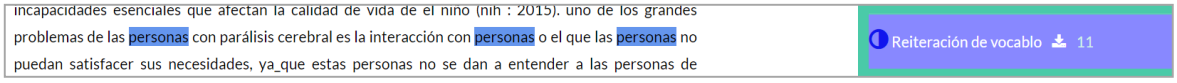


Figura 81 Errores de reiteración de vocablos analizados por la aplicación Web.

Errores de aberraciones lingüísticas

En la Figura 82, se muestra el tipo de errores llamados “aberraciones lingüísticas” que se detectan en la interfaz Web.

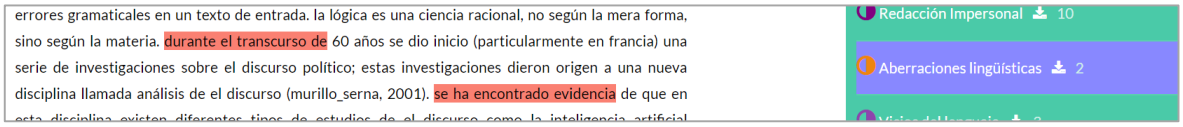


Figura 82 Errores de aberraciones lingüísticas analizados por la aplicación Web.

Referencias

[1] J. J. Comparán Rizo, C. G. Amezcua Rosales, A. Arriaga González, and G. Bañuelos Valera, *Lengua Española 3*. 2007.

[2] A. Corbacho, "Textos , Tipos De Texto Y Textos Especializados," *Artículo Científico*, vol. 3, pp. 77–90, 2006.

[3] E. I. Moyano, "Una clasificación de géneros científicos," *XIX Congr. AESLA*, pp. 1–9, 2001.

[4] C. Prensa, "México, entre los 10 países con más graduados en ciencias." 2016.

[5] C. Culebra y Vives, "Taller de ortografía y redacción básicas," 2004.

[6] Medidas de evaluación de la escritura, realizadas por La Dra. Margarita Peón Zapata y la Mtra. Sara Rivera ambas de la Dirección de Pruebas y Medición del INNE.

[7] Rae, "Diccionario de la Lengua Española," *Mod. Lang. J.*, vol. 11, no. 6, p. 391, 2001.

[8] "DLE_ cacofonía - Diccionario de la lengua española - Edición del Tricentenario." [Online].

[9] E. A. Llorach, "Gramática de la lengua española," p. 508, 2000.

[10] M. de Hoyos González, "Una variedad en el habla coloquial: la jerga'chelí'," *Cauce Rev. Filol. y su didáctica*, no. 4, pp. 31–42, 1981

[11] N. Vicente, "Las variedades de la lengua." pp. 1–15.

[12] RAE, "Real Academia Española," *Real Academia Española, Asociación de Academias de la Lengua Española. Diccionario de la lengua española, 23.ª ed., Edición del Tricentenario, [en línea]. Madrid: España, 2014*. 2014.

[13] Real Academia Española, "Términos lingüísticos | Real Academia Española." 2015.

[14] J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró, "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library," *Proc. 5th Int. Conf. Lang. Resour. Eval. Lr.*, no. January, pp. 48–55, 2006.

[15] V. Daudaraviˇ and H. Ford, "Automated Evaluation of Scientific Writing : AESW Shared Task Proposal," pp. 56–63, 2015.

[16] D. Naber, "A Rule-Based Style and Grammar Checker," pp. 1–77, 2003.

[17] M. Mozgovoy, "Dependency-based rules for grammar checking with LanguageTool," *2011 Fed. Conf. Comput. Sci. Inf. Syst.*, pp. 209–212, 2011.

[18] V. Daudaravicius, R. E. Banchs, E. Volodina, and C. Napoles, "A Report on the Automatic Evaluation of Scientific Writing Shared Task," *NAACL BEA11 Work.*, pp. 53–62, 2016.

[19] J. T. Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, *Shared Task on Grammatical Error Correction*. 2014.

[20] V. Daudaravicius, R. E. Banchs, E. Volodina, and C. Napoles, "A Report on the Automatic Evaluation of Scientific Writing Shared Task," *NAACL BEA11 Work.*, pp. 53–62, 2016.

[21] T. Turunen, "Introduction to Scientific Writing Assistant (SWAN) – Tool for Evaluating the Quality of Scientific Manuscripts," no. May, 2013.

[22] L. G. Vidales, "Tesis (en desarrollo)Dominio de características lingüísticas para redactar textos técnicos en estudiantes del área de Computación," 2016.

[23] D. A. Maura, "Deficiencias frecuentes en la redacción de textos científicos."

[24] C. R. Orlando, "Errores frecuentes en el uso de preposiciones." 2016.

[25] U. A. de Barcelona, "Corrección, estilo y variaciones de la lengua española,".

[26] R. A. E. de la Lengua and R. A. E. de la Lengua, "Diccionario Panhispánico de Dudas."

[27] DLE_ barbarismo - Diccionario de la lengua española - Edición del Tricentenario.

[28] P. José, ESTUDIOS de LINGÜÍSTICA "Cambios fonéticos esporádicos: metaplasmos, vulgarismos o licencias fonológicas." 2002.

[29] "Estilo y redacción del trabajo final." Universidad Católica Andrés Bello, Caracas, Venezuela, pp. 25–27, 2008.

[30] E. B. Escudero, «Evaluación de la competencia de expresión escrita argumentativa de estudiantes universitarios,» *revista de la educación superior,* pp. pp. 9 - 39, 2013.

[31] J. A. Moreiro, «Criterios e indicadores para evaluar la calidad del análisis documental de contenido,» Ci. Inf., Brasília, v. 31, pp. 53-60, 2002.

[32] Instituto Nacional para la Evaluación de la Educación [INNE] (2011). La Educación Media Superior en México.México, D.F.: INNE.

[33] Lavy, V. (2010). Do Differences in School's Instruction Time Explain International Achievement Gaps in Maths, Science and Language? Evidence from Developed and Developing Countries. Londres: Centre for the Economics of Education. Recuperado de: http://cee.lse.ac.uk/ceedps/ceedp118.pdf.

[34] Rosales, P. y Vázquez, A. (2006). Escribir y aprender en la Universidad. Análisis de textos académicos de los estudiantes y su relación con el cambio cognitivo. signoEseña, 16 (diciembre) 34-47.

[35] Secretaría de Educación Pública [SEP] (2012). Evaluación Nacional del Logro Académico en Centros Escolares (ENLACE)-Educación Media Superior. http://201.175.44.204/Enlace/Resultados2012/MediaSuperior2012/R12msOtrosCriteriosConsulta.aspx.

[36] ABAD GARCÍA, F. Investigación evaluativa en documentación. Valencia : Universitat de Valencia, 1997. p. 130-132

[37] AENOR (Madrid). UNE 50-121-91: métodos para el análisis de documentos, determinación de sua contenido y selección de los términos de indización. Madrid, 1991.

[38] BLAIR, D. Language and representation in information retrieval. Amsterdam : Elsevier, 1990.

[39] CHAUMIER, J. Análisis y lenguajes documentales. Barcelona : Mitre, 1986.

[40] DENIS, S. et al. Liability in the provision of information services. Brussels : EUSIDIC, 1990.

[41] ELLIS, D. The effectiveness of information retrieval systems: the need for improved explanatory frameworks. Social Science Information Studies, n. 4, p. 261-272, 1984.

[42] EISENBERG, M. Measuring relevance judgments. Information Processing and Management, n. 24, p. 373-389, 1988.

[43] FOLSTER, M. B. A study of the use of information sources by social science researchers, Journal of the Academic Librarianship, n. 1, p. 7- 11, 1989.

[44] GRIFFITHS, J.; KING, D. A manual on the evaluation of information centers and services. Neuilly-sur-Seine : North Atlantic Treaty Organization, 1991.

[45] INFORMATION MARKET OBSERVATORY. The quality of electronic Information products and services. Luxembourg, 1995. (Working Paper, 95/4).

[46] LABOIRE, T.; HALPEIN, M.; WHITE, H. Library and information science abstracting and indexing services: coverage, overlap and context. Library and Information Science Abstracts, n. 7, p. 183-195, 1985.

[47] UNESCO (Paris). Principes directeurs pour l'évaluation des systèmes et services d'information. Paris : UNESCO, 1978.

[48] ROLLING, L. Indexing consistency, quality and efficiency. Information Processing and Management, v. 17, p. 69-76, 1981.