PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

HERRAMIENTA DE ANÁLISIS Y CLASIFICACIÓN DE COMPLEJIDAD DE TEXTOS EN ESPAÑOL

Tesis para optar el Título de **Ingeniero Informático,** que presentan los bachilleres:

Walter Perez Urcia
André Raúl Quispesaravia Ildefonso

ASESOR: Fernando Alva Manchego

Lima, enero de 2015

RESUMEN

La selección de textos es una de las responsabilidades principales de los profesores dentro del planeamiento del orden de exposición a textos para sus alumnos. Debido a la gran cantidad de textos disponibles y la diversidad de géneros y temas, esta tarea demanda mucho tiempo y está ligada a aspectos subjetivos del evaluador. Esto es un problema, sobre el cual proponemos una alternativa de automatización.

Se toma como proyecto la implementación de una herramienta de análisis y clasificación de complejidad de tetos en español. Con ello se busca brindar una alternativa automatizada al problema de escalabilidad en complejidad de textos. Esto se refiere a la necesidad de tener un orden de textos evaluados por complejidad.

Para ello evaluamos la complejidad utilizando las métricas de Coh-Metrix adaptadas al español. Este conjunto de métricas evalúa textos en inglés de acuerdo a los estudios de coherencia y cohesión que los autores de Coh-Metrix desarrollaron.

En base a esta adatapación de Coh-Metrix se desarrolló un clasificador basado en técnicas de aprenizaje de máquina y un conjunto de textos en español conformado por fábulas y cuentos previamente clasificados. Finalmente estos componentes fueron unidos en una herraamienta web para la accesibilidad pública de la herramienta desarrollada.

Tabla de contenido

Índice de Figuras

Índice de tablas

CAPÍTULO 1

1 Introducción

La selección de textos es una de las responsabilidades principales de los profesores dentro del planeamiento del orden de exposición a textos para sus alumnos. Debido a la gran cantidad de textos disponibles y la diversidad de géneros y temas, esta tarea demanda de mucho tiempo y está ligada a aspectos subjetivos del evaluador [Fisher, et al, 2012]. Esto es un problema, sobre el cual proponemos una alternativa de automatización.

Se toma como proyecto la implementación de una herramienta de análisis y clasificación de complejidad de textos en español. Con ello se busca brindar una alternativa automatizada al problema de escalabilidad en complejidad de textos. Esto se refiere a la necesidad de tener un orden de textos evaluados por complejidad.

Para ello evaluamos la complejidad utilizando las métricas de Coh-Metrix. [Graesser, et al, 2011]. adaptadas al español. Este conjunto de métricas evalúa textos en inglés de acuerdo a los estudios de coherencia y cohesión que los autores de Coh-Metrix desarrollaron.

En base a esta adaptación de Coh-Metrix se desarrolló un clasificador basado en técnicas de aprendizaje de máquina y un conjunto de textos en español conformado por fabulas y cuentos previamente clasificados. Finalmente estos componentes fueron unidos en una herramienta web para la accesibilidad pública de la herramienta desarrollada.

En este capítulo se describirá la problemática en la que se basa el proyecto de fin de carrera, los conceptos base que se cubrieron en el desarrollo del proyecto y los trabajos relacionados a este campo.

2 Problemática

La lectura es una de las herramientas de mayor utilidad y gran importancia en el aprendizaje de muchas áreas de conocimiento. Esto se debe a que la transmisión del conocimiento en la humanidad se da por medios orales y escritos, y es la lectura la que nos permite acceder a esta información [OECD, 2012].

Sin embargo, se debe tener en cuenta que no todos los textos tienen las mismas características y que el nivel de comprensión lectora necesario para interiorizar esta información es también variable. Se tiene que identificar el nivel de lenguaje, el sentido, comprender las metáforas y argumentos literarios utilizados dentro del texto. También se necesario encontrar el contexto histórico, se necesita de hacer un análisis critico del texto. Comprender satisfactoriamente un texto depende del conocimiento previo, los mecanismos de inferencia, la habilidad del lector y su contexto sociocultural. [Graesser, et al, 2011].

Los estudiantes de nivel escolar se encuentran con textos de diversas fuentes que deben leer para adquirir conocimiento acerca de algún tema. Sin embargo, no siempre se llega a comprender estos textos, sobre todo en las etapas iniciales de aprendizaje como el colegio, ya que la exposición a textos complejos no es parte del proceso de enseñanza y esta exposición solo es parte de un proceso largo [Fisher, et al, 2012].

Para analizar la capacidad de comprensión lectora de los alumnos, se han realizado exámenes (PISA) cuyos resultados muestran que este aspecto es deficiente dentro del Perú y también en comparación con la media obtenida por alumnos de otros países.

A nivel mundial, el examen PISA, que se lleva a cabo aproximadamente cada 3 años desde 1997, se concentra en examinar las estrategias de aprendizaje de los estudiantes, sus habilidades en áreas en las que intervienen múltiples disciplinas y distintos temas de interés, concentrándose en lectura, matemáticas y ciencias. En su informe más reciente (2012) se presta especial atención a la habilidad de lectura en diversos dominios de conocimiento, contextos y situaciones; se evaluó la capacidad de leer para aprender más que el aprendizaje de la lectura. Los resultados de este informe muestran que el Perú se encuentra en la última posición de 64 países analizados en la prueba en general y su puntuación de comprensión lectora está por debajo del promedio mundial [OECD, 2012].

De la misma forma, a nivel nacional, en el último censo estudiantil tomado a alumnos de segundo y cuarto grado de primaria, se puede observar que un 15.8% tiene una comprensión lectora muy pobre, pudiendo apenas responder preguntas básicas sobre lo que leyó. Se podría asumir que el otro porcentaje de alumnos (84.2%) tiene una buena comprensión, pero este no es el caso dado que el 51.3% del total de alumnos tiene una comprensión lectora en la que solo pueden identificar información de textos cortos, pero sin sacar conclusiones sobre estos. Es decir, apenas un 32.9% logra analizar textos largos y llegar a conclusiones a partir de la información que identifica en ellos [MINEDU, 2013].

Se puede notar que más de la mitad de alumnos censados tienen problemas para poder entender claramente y obtener conclusiones de lo que lee. Estas cifras son incluso más preocupantes si son analizadas a nivel regional, siendo mayores en las zonas rurales del país, donde puede llegar hasta casi un 90% del total con problemas de comprensión [MINEDU, 2013].

Se observa que el problema es principalmente la falta de comprensión de textos desde temprana edad. Las posibles razones por las que ocurre esto son: la pobreza de vocabulario, escasos conocimientos del tema, problemas de memoria, textos inadecuados para la edad o que sean de difícil comprensión, entre otros [Fisher et al., 2012]. Esta situación afecta de manera directa al futuro de los estudiantes, ya que la capacidad de comprensión de lectura es fundamental para el aprendizaje y la expectativa que se tiene del nivel de esta capacidad por parte de los empleadores crece con el nivel de estudios que se tiene [SHRM, 2006].

Para mejorar esta deficiencia en la capacidad lectora, se requiere de una mayor exposición a textos y de escalar la complejidad de estos de manera gradual. Es por ello que el encargado de esta actividad tiene que conocer cómo clasificar los textos que se proponen a los estudiantes. Esto debido a que deben tener una complejidad adecuada a su nivel de comprensión lectora y se tiene que tener en cuenta que los estudiantes están tanto aprendiendo a leer como leyendo para aprender [Graesser, et al, 2011].

La selección de textos es una de las responsabilidades principales de los profesores dentro del planeamiento del orden de exposición a textos. Debido a la gran cantidad de textos disponibles y la diversidad de géneros y temas, esta tarea demanda de mucho tiempo y está ligada a aspectos subjetivos del evaluador [Fisher, et al, 2012].

Es por ello que para la evaluación de textos y su escalamiento en dificultad, se definen varias características para identificar su legibilidad y complejidad. Estas son su estructura, coherencia, unidad y cuán apropiado es para la audiencia a la que se dirige. Estas características deben incluir tanto atributos cualitativos como escalas cuantitativas por las que tienen que medirse. La medición de estas características no siempre es posible, ya que a nivel cualitativo se requiere del análisis por parte de un lector humano, pero también existen herramientas que ayudan a esta tarea mediante el uso de métricas para el caso del análisis cuantitativo [Fisher et. al, 2012].

Estas herramientas para la extracción de métricas dentro del análisis cuantitativo están directamente relacionadas con el lenguaje que analizan y son útiles en varios estudios realizados en clasificación de textos, legibilidad, simplificación, extracción de significados, etc. [Graesser, et al 2011]. Para el idioma inglés, existe diversidad de herramientas que cumplen esta función e implementan métricas básicas y avanzadas que pueden ser adaptadas al español.

Para el apoyo al análisis automatizado de textos en el proceso de escalamiento y selección el presente proyecto plantea el desarrollo de una herramienta que apoye a la clasificación de complejidad textual de forma automatizada. Esto en base al análisis de textos utilizando datos numéricos (también conocidos como indicadores o métricas lingüísticas) para la clasificación de su complejidad. Para esto, se planea adaptar para el idioma español la herramienta Coh-Metrix. Esta herramienta analiza textos en varios niveles del lenguaje, extrae métricas que evalúan características de textos. Coh-Metrix ayudó a varios estudios de procesamiento de texto en inglés [McNamara, 2010] y también fue adaptada para el idioma portugués, como parte de proyectos de simplificación textual [Scarton, 2010].

Con esta adaptación se hará posible la extracción de métricas de los textos a estudiar, los cuales extraerán propiedades básicas de los textos por medio de conteo y algunas propiedades más avanzadas como cohesión y coherencia. Con estas métricas se podrá analizar detalladamente los textos y también desarrollar un clasificador de complejidad textual por medio de la aplicación de técnicas de aprendizaje de máquina.

3 Marco teórico

A continuación se explican conceptos básicos utilizados en el análisis de la problemática relacionada con la importancia de la lectura y la complejidad textual. También se describen algunos métodos de procesamiento de lenguaje natural y aprendizaje de máquina que se utilizan en investigaciones de complejidad y análisis de texto plano.

3.1 Marco conceptual

En las siguientes secciones se explicarán conceptos relacionados con el problema planteado, el proceso de análisis de complejidad de textos y las técnicas y términos de aprendizaje de máquina utilizados en las investigaciones revisadas y en la propuesta de solución.

3.1.1 Conceptos relacionados al problema

En la problemática se abarcaron los temas de:

- **Comprensión lectora**

  Se refiere a la capacidad por parte del lector de poder extraer ideas del texto, relacionarlas y poder sacar conclusiones a partir de lo leído. Esta puede ser medida por diferentes tipos de preguntas, desde las más básicas, como el título y tema central, hasta más complejas como a qué conclusión podrías llegar teniendo en cuenta lo que menciona el autor o qué relación tienen el primer y último punto que se menciona en el texto. A cada pregunta y cada alternativa mostrada se le otorga puntajes y estos finalmente son procesados para obtener un indicador de comprensión. En la Figura 1 se muestra un ejemplo de pregunta tomada en el examen PISA (examen a nivel mundial para la medición de comprensión lectora y matemáticas) en la que se pide una pequeña descripción del texto [PISA, 2007].

Figura 1 Pregunta de ejemplo, Examen PISA.[1]

- **Términos generales**

  Se usará el término **legibilidad** para referirse a los términos en inglés *readability* y *legibility*. En la lengua inglesa estos términos son diferenciados uno del otro: el primero tiene que ver con la forma en que se usan y combinan las palabras para formar frases y párrafos completos, mientras que el segundo se refiere más a aspectos viso-espaciales del texto. Algunos traducen el primer término como lecturabilidad o comprensibilidad; sin embargo, a lo largo de este proyecto se usará legibilidad para referirse a ambos términos siguiendo la convención establecida por [Legibilidad, 2007].

  Cohesión y **coherencia** son conceptos muy relacionados, pero bien diferenciados. Cohesión está relacionada con las características del texto que ayudan al lector a conectar ideas y poder hacer una representación mental del texto completo. Por otro lado, coherencia es el conjunto de características que le dan sentido a la representación mental que el lector realiza sobre el texto [Arthur, 2004].

[1] **Fuente: http://pisa-sq.acer.edu.au/showQuestion.php?testId=2292&questionId=12**

Por último, una **métrica** es un indicador que expresa numéricamente alguna característica de un texto como propiedades morfológicas, sintácticas, de cohesión, entre otras. Pueden ser desde simples contadores de palabras o sílabas, hasta métricas más complejas de legibilidad de textos, como es el caso de la métrica de legibilidad Fernandez-Huertas [Fernandez, 1959]. Estas métricas pueden ayudar a tener una idea más clara sobre los datos del texto y no tan solo una idea descriptiva. Son usados principalmente para el análisis de legibilidad de textos.

- **Ambigüedad**

  Se dice que una idea es ambigua cuando existe más de una estructura lingüística para ella; es decir, que puede tomarse de diferentes formas dependiendo de cómo esta esté redactada [Jurafsky, 2007]. Considere los siguientes ejemplos:

Tabla 1 Ejemplos de ambigüedad

Oración	Ambigüedad
Lourdes no quiere a su tía porque es muy envidiosa.	No queda claro si la persona envidiosa es Lourdes o su tía
Estuve esperándote en el banco.	No se sabe si estuvo sentado en un banco o en una institución financiera
El Real Madrid le ganó al Barcelona en su campo.	No se puede determinar con exactitud de quién es el campo donde se jugó

Para el análisis de la complejidad textual se hizo uso de:

- **Componentes de la complejidad textual**

  Se define complejidad textual como tres componentes interrelacionados, como se puede observar en la Figura 2. El componente **cualitativo**, que mide la tentativa de un lector humano, como el significado, el propósito, las convenciones utilizadas; el componente **cuantitativo**, que da factores cuantitativos, como longitud de las palabras, oraciones, cohesión en el texto; y **el lector y la tarea**, que tiene en cuenta las particularidades del lector, como su domino del área de conocimiento del texto, motivación y experiencias.

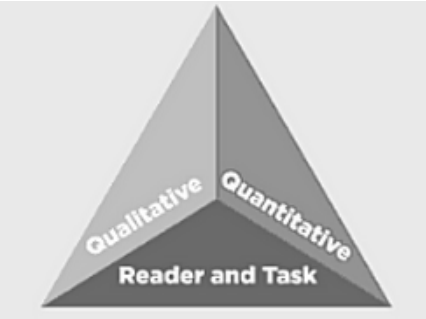


Figura 2 Representación de los componentes de la complejidad. Fuente: [Fisher, 2012]

3.1.2 Conceptos relacionados a la propuesta de solución

Se propone como solución la adaptación de una herramienta de extracción de métricas de un texto, la cual analizará las oraciones de los textos a evaluar y extraerá información acerca de sus propiedades sintácticas y semánticas, así como estadísticas sobre las palabras utilizadas. Después, con estas métricas se usarán técnicas de aprendizaje de máquina para desarrollar un clasificador de textos según su complejidad.

A continuación, se describen los conceptos básicos utilizados en la revisión bibliográfica sobre extracción de métricas a partir de texto plano.

- **Análisis de una oración**

  Una oración se puede analizar según tres criterios: morfología, sintaxis y semántica.
  La **morfología** está relacionada a la función que tiene cada palabra en una oración o frase. La **sintaxis** tiene que ver con la estructura o la forma en que se organizan las palabras en una oración. Por último, la **semántica** está enfocada al significado o sentido que tiene la oración [Jurafsky, 2004]. Por ejemplo:

  - José compró una bicicleta el día de ayer.

  Según la morfología, “compró” sería el verbo principal. Según la sintaxis de la oración, “compró una bicicleta el día de ayer” sería una frase verbal y el significado (semántica) de la oración sería que José adquirió una bicicleta ayer.

- **Análisis Morfológico**

  Para poder entender este tipo de análisis se debe tener en cuenta otros tres tipos de análisis: *stemming, lemmatization* y *tokenization*. El primero tiene que ver con la identificación de la raíz principal de una palabra; por ejemplo, libro y librería tienen una raíz en común libr-. El segundo término está relacionado con el anterior, pero más enfocado a las posibles conjugaciones de estos; por ejemplo, cantó, cantamos y cantarás, provienen de un mismo verbo, cantar. El último de los términos implica delimitar y separar las palabras del texto según su función en él: modificadores, objetos, sustantivos, entre otros [Jurafsky, 2007].

Para el análisis de textos y métodos de clasificación se utilizan técnicas de aprendizaje de máquina y modelos estadísticos del lenguaje.

- **Modelos estadísticos del lenguaje**

  Los modelos estadísticos del lenguaje son usados en procesamiento de lenguaje natural para asignar una probabilidad a una secuencia de palabras considerando la probabilidad de cada palabra en su historial. Los modelos explorados son de *n-gramas*, que basan su funcionamiento en la propiedad de que una secuencia de palabras cumple la propiedad de Markov; es decir, la probabilidad de una palabra es condicionada por las n palabras anteriores [Jurafsky et al, 2007].

El modelo de n-gramas soporta dependencia total del historial, pero al ser esto poco práctico, los modelos usados son los de 1-gramas, 2-gramas y 3-gramas. Para el modelo, usualmente las oraciones son tratadas independientemente [Petersen, 2007].

- **Árbol de análisis sintáctico**

  Es un árbol con raíz que representa un conjunto de elementos y un conjunto de etiquetas, donde todos los nodos tienen paridad mayor o igual a dos. Las etiquetas del árbol son las características del lenguaje usado y las hojas los elementos del conjunto [Chriswell & Hodges, 2007]. Por ejemplo, en la Figura 3 se presenta el árbol de sintaxis de la oración en inglés "workers dumped stacks into a bin", de la cual se identifica como Sintagma nominal "workers" y Sintagma verbal, "dumped sacks into a bin".

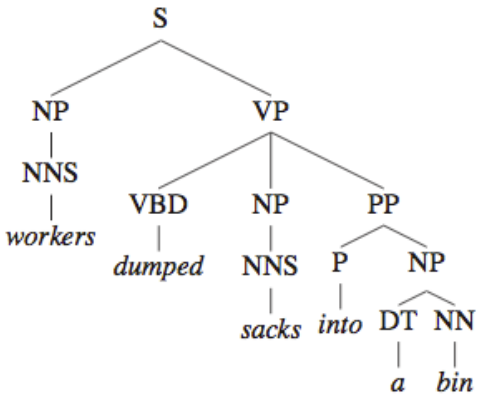


Figura 3 Árbol de análisis sintáctico de una oración en inglés. Fuente: [Chiswell & Hodges, 2007]

De estos modelos estadísticos e información extraída de los árboles de análisis sintáctico se eligen algunas características (“*features*”) que son utilizadas en el proceso de aprendizaje de máquina para entrenar un clasificador.

- **Aprendizaje de máquina**

  Se denomina aprendizaje de máquina al uso de modelos basados en aprendizaje estadístico para el desarrollo de sistemas que puedan adaptarse de acuerdo a este modelo y al procesamiento de datos históricos. Es decir, aprender a partir de datos. Este aprendizaje tiene como objetivo el predecir alguna medida, que puede ser cuantitativa o categórica. La predicción es basada en características que se extraen del escenario que se está observando [Hastie et al, 2008].

  Una definición más rigurosa propuesta por Mitchell señala:

  > *Se entiende por aprendizaje de máquina de un programa que aprende de la **experiencia E** con respecto a alguna clase de **tarea T** y medida de **rendimiento P**, si el rendimiento en la tarea **T**, que es medido por **P**, mejora con la experiencia **E*** [Mitchell, 1997].

Para realizar este aprendizaje en base a la experiencia, inicialmente se tiene un conjunto de datos de entrenamiento, en el que se observan las mediciones de características que se desean extraer. Estas características pre-procesan el escenario que se desea estudiar, tratando de transformarlo en algún espacio en donde se espera que el problema disminuya en complejidad. Este pre-procesamiento toma el nombre de extracción de características [Bishop, 2006].

Se debe tener en cuenta que si bien todas las características que se pueden proponer sobre la tarea evaluada son válidas y contienen información, no todos estos factores pueden ser relevantes. Es por ello que se deben tener en cuenta factores como la entropía, la dimensión del problema y el tipo de problema. Siendo así, se pasa por un proceso de selección de características (que puede ser manual o automatizado) en el que se filtren y elija un conjunto que disminuya la dimensión del problema y brinde la información sobre el escenario con la menor redundancia [Bishop, 2006].

Después de observar y elegir estas mediciones, se pasa a un proceso de pruebas de hipótesis, en el cual se elige un modelo de aprendizaje para la tarea en base a los datos de prueba y que pueda adaptarse. Finalmente, se comprueba la validez de este modelo, contrastándolo con un segundo conjunto de datos en los cuales se pueda comparar la predicción con los resultados reales en base a la medida de rendimiento elegida [Hastie et al, 2008; Mitchell, 1997].

Se puede notar que la tarea es el factor que define el problema que se quiere resolver. Es por ello que dependiendo de las características de la experiencia, tarea y medida de rendimiento utilizada se pueden clasificar a las tareas de aprendizaje de máquinas en aprendizaje supervisado, aprendizaje no supervisado y reforzamiento de aprendizaje.

Se entienden por problemas de ***aprendizaje supervisado*** cuando los datos de entrenamiento contienen ejemplos de características y sus resultados correspondientes; cuando el objetivo del problema es asignarle un resultado por medio de la inferencia basada en la experiencia. Cuando este resultado a asignar pertenece a un conjunto discreto de categorías la tarea es llamada ***Clasificación,*** en caso el resultado a asignar pertenece a uno o más datos continuos la tarea es llamada ***Regresión.***

En caso el objetivo no es asignarle un resultado a un problema, es decir se cuenta solamente con datos pero sin sus valores de resultado, la tarea pertenece a ***aprendizaje no supervisado.*** El objetivo en este caso es el descubrir la estructura interna de los datos bajo el conjunto de características a analizar. Si se busca encontrar grupos de datos similares, la tarea es de ***Clustering (Agrupamiento),*** en caso se requiera solo encontrar cual es la distribución de los datos en el espacio evaluado se refiere a ***Estimación de densidad*** y si se requiere brindar una visualización de datos de altas dimensiones a dos o tres dimensiones, se trata de ***Visualización.***

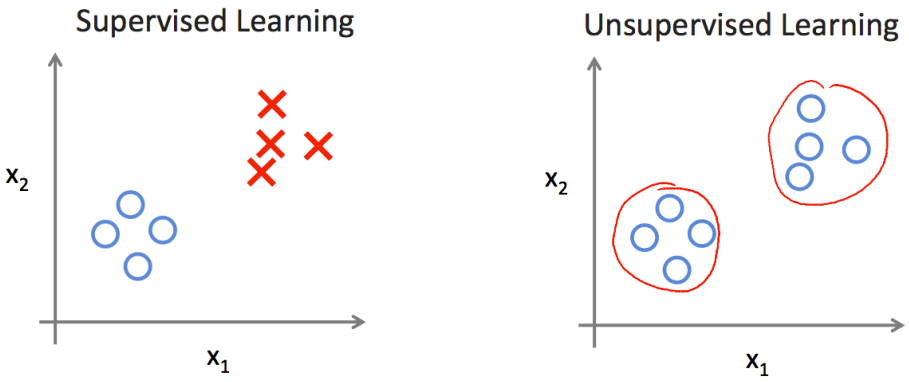


Figura 4 Aprendizaje supervisado (izquierda) y no supervisado (derecha)
Fuente: [Ng, 2013]

Como se puede observar el la Figura 4[2], la principal diferencia entre el aprendizaje supervisado (izquierda) y no supervisado se encuentra en el conocimiento de las características de los datos en estudio. Mientras que en el supervisado se tienen características de estructura definidas, en el no supervisado se intenta asignar una estructura a los datos estudiados.

En caso se requiera encontrar una acción que maximice la ganancia en alguna situación, se está refiriendo a ***Reforzamiento de aprendizaje***. En este caso no se toma como base a datos que contenga salidas que beneficien a mi ganancia, sino que el algoritmo descubre estas acciones por medio de prueba y error [Bishop, 2006].

- **Procesamiento de lenguaje natural apoyado por métodos estadísticos**

  Los seres humanos utilizan el lenguaje para comunicarse. Este es un conjunto de reglas dependientes del idioma hablado que brindan una base con la cual los individuos pueden compartir sus ideas. La ciencia lingüística se encarga de darle formalidad a estas reglas para que se tenga una interpretación única a la estructura del lenguaje. Pero el mayor problema con ello es que estas reglas no son siempre respetadas por los hablantes y siempre existe una evolución que hace que estas reglas se adecuen al uso que se les da y no que los hablantes se adecuen a las reglas.

  Considerando lo anterior, se necesita explorar métodos que ayuden a interpretar el lenguaje humano y que tomen en cuenta más que reglas estáticas; por ejemplo, que identifique patrones que ocurren en el lenguaje. Es por ello que el objetivo del procesamiento de lenguaje natural (PLN) es desarrollar modelos y métodos que permitan a las computadoras desarrollar tareas útiles relacionadas con el lenguaje humano [Jurasfsky, 2006].

[2] Extraída de https://d396qusza40orc.cloudfront.net/ml/docs%2Fslides%2FLecture1.pdf

La dificultad en PLN está en la ambigüedad implícita en todo texto. Decimos que un texto es ambiguo si hay muchas alternativas para la construcción de estructuras lingüísticas que pueden ser construidas para la misma entrada. Esta ambigüedad es fácilmente identificada e interiorizada por un humano pero lo que se busca es realizar modelos que también sean capaces de lo mismo. Es por ello que se adopta una metodología de trabajo para resolver estos problemas automatizando el aprendizaje léxico y estructural desde un conjunto de textos (*Corpora*).

Este aprendizaje tiene mucha similitud con el de aprendizaje de máquinas. Se toman como características a extraer del *corpora* las relaciones que se tienen entre palabras, los grupos de palabras que tienden a estar juntas, etc., para luego aplicar modelos estadísticos que confirmen sus profundas relaciones semánticas, sintácticas, etc. Mediante esta metodología se reduce el esfuerzo humano en la producción de sistemas de PLN que estarían solamente basados en reglas estáticas.

Los problemas que son tratados en PLN se pueden clasificar en alineación estadística y traducción, *clustering*, recuperación de información y clasificación de textos.

En ***alineación estadística y traducción***, se ven problemas relacionados a la traducción automática de texto o voz de algún lenguaje a otro; en ***clustering*** se busca partir a un conjunto de textos en grupos en los cuales solo existan similares bajo el concepto de igualdad utilizado; en ***recuperación de información*** se busca desarrollar algoritmos y modelos que recuperen información de documentos, en especial textos; finalmente, en ***clasificación de textos*** se busca asignarle clases o etiquetas a los textos evaluados [Manning, 1999].

4 Estado del arte

Se describirá el estado de las herramientas e investigaciones que estén relacionadas con el problema de análisis de complejidad de textos. Como soluciones aproximadas se mencionarán las alternativas en otros idiomas distintos al español. Hay que mencionar que los problemas del área de procesamiento de lenguaje natural no tienen solución exacta, todas las investigaciones y herramientas mencionadas proponen un modelo como alternativa de solución al problema.

4.1 Formas aproximadas de resolver el problema.

Como método de solución del análisis de complejidad de textos se tiene a la extracción de métricas del texto por medio de una herramienta, clasificación manual de un conjunto de datos y luego la aplicación de las fases de los algoritmos de aprendizaje de máquina supervisados.

A continuación se presentan algunas alternativas de clasificación e investigación de complejidad de textos por medio del uso de la metodología anteriormente descrita.

- **Ordenamiento de texto por legibilidad**

  En [Tanaka, 2002] se describe un método de clasificación de textos en base a su complejidad basado en el uso de un operador binario de orden. El objetivo de esta investigación fue el de brindar un nuevo método de clasificación de complejidad textual.

El problema que plantean es que se necesita una gran cantidad de texto previamente clasificado para poder entrenar y obtener buenos resultados con algoritmos de clasificación basados en aprendizaje de máquina y clasificación. Este problema se agrava porque no se pueden utilizar algunos textos clasificados a causa de sus patentes y restricciones de uso. También hace notar que las métricas usadas por los métodos que utilizan la valorización de la complejidad del texto por medio de una fórmula y extrapolación de su valor son muy básicas y dependientes del lenguaje estudiado.

Es por estas razones que los autores optan por no usar clasificaciones pre establecidas para su método de clasificación de complejidad textual; por el contrario, usan un comparador binario que pueda determinar la relación de dificultad entre los dos textos comparados. Por medio del uso de este operador binario, se libera de la restricción de tener texto clasificado previamente para varias clases y le queda el trabajo de determinar un entrenamiento para este operador. Este entrenamiento necesita de pares de textos en los que se determine de antemano cuál texto es más complejo. Para el entrenamiento del clasificador se utiliza un SVM (Support Vector Machine[3]).

Con ello puede tener como clasificación un ordenamiento de textos por complejidad. El estudio no plantea sustituir a clasificadores en donde se divide claramente a los textos en niveles específicos (textos para 1er grado, 2do grado, etc.), sino plantea establecer un orden de complejidad entre los textos evaluados. En esta estructura ordenada pueden hacer consultas de cuál sería la complejidad de un texto en un conjunto de textos previamente ordenados. Las características que toma en cuenta el operador binario para las comparaciones del texto son el uso de vocabulario de palabas frecuentes.

Para llegar a este ordenamiento, los autores usan el algoritmo de inserción. Esto lo hacen ya que, por su metodología, se propone agregar un texto a la vez al conjunto ordenado, y para confirmar que se encuentra en el lugar correcto, se compara al texto insertado con los textos ya ordenados que se encuentren a una distancia configurable, como puede observarse en la Figura 5.

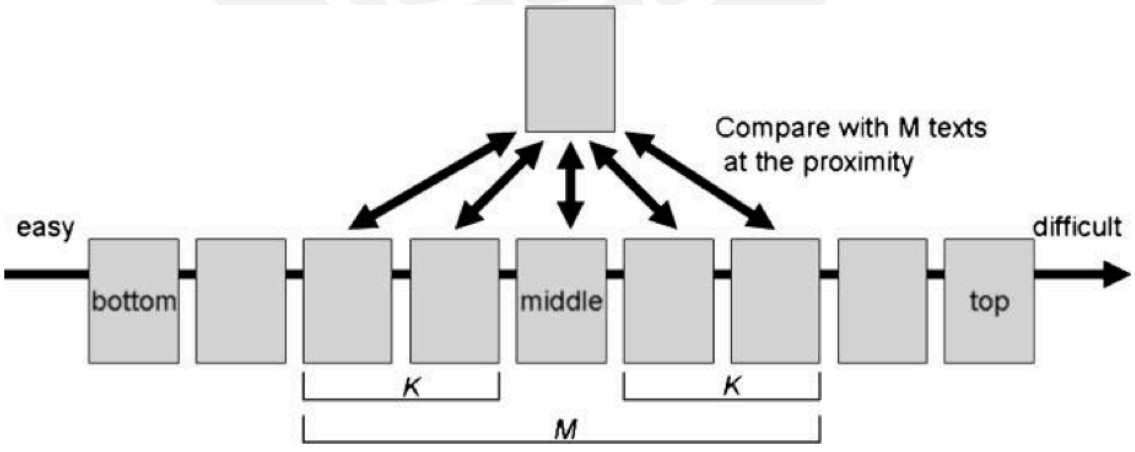


Figura 5 Diagrama de comparación y visualización del orden. Fuente: [Tanaka, 2002]

[3] SVM (Support Vector Machine): modelo de aprendizaje supervisado en el que la determinación de los parámetros está dado por optimización en funciones convexas [Bishop, 2006].

Para la consulta de la complejidad de un texto, se utiliza una búsqueda binaria, comparando con los anteriores y siguientes en orden en cada iteración.

Los resultados de este estudio muestran que este ordenamiento de textos y su robusta comparación presentan una mayor correlación entre sus resultados. Los autores de este artículo no pretenden proponer su método como el mejor, ya que si los demás algoritmos de clasificación llegan a tener mayor cantidad de textos de entrenamiento pueden llegar a tener mejores resultados [Tanaka, 2002].

- **Evaluación de nivel de lectura utilizando máquinas de vectores de soporte y modelos estadísticos de lenguaje**

  En [Petersen, 2008], los autores proponen un modelo de clasificación de complejidad textual basado en árboles de análisis sintáctico y modelos estadísticos del lenguaje como características a evaluar por el método de máquinas de vectores de soporte (SVM). El contexto de su problema es el crecimiento de la enseñanza multilingüe y propone el desarrollo de una herramienta para la clasificación de complejidad de textos para esta área.

  Esta investigación menciona y señala los problemas de los métodos estadísticos basados en el conteo de características del texto, los cuales son muy dependientes del lenguaje y pueden ser mejorados con técnicas de aprendizaje de máquinas. Es por ello que los autores proponen el estudio de estos algoritmos basados en aprendizaje de máquinas para analizar datos clasificados en 4 niveles de dificultad: ediciones de 2do a 5to grado del Weekly Reader[4].

  Como modelo estadístico de lenguaje se eligió a los n-gramas, que modelan a un orden de palabras como un proceso de decisión de Markov. Las propiedades de los árboles de análisis sintáctico utilizadas fueron: longitud promedio, número de frases sustantivas, frases verbo y SBARS. Todas estas características fueron filtradas por un proceso de selección en el que se removieron propiedades redundantes y de poca correlación [Scharm, 2005].

  Este trabajo tiene una segunda parte, en la que se extiende el uso de las SVM para que pueda aceptar nuevos datos clasificados en categorías distintas a las que fue entrenado el SVM y que sea capaz de reconocer estas nuevas categorías. También experimentan con regresión lineal y exponen el problema de la subjetividad de la complejidad textual, ya que para ciertos campos se usan diferentes criterios de complejidad y esto puede ser remediado agregando datos de entrenamiento clasificados manualmente con las especificaciones de los usuarios [Petersen, 2008].

- **Análisis lingüístico de textos simplificados y auténticos**

  En este análisis se estudia la complejidad textual de textos auténticos (textos que son escritos en inglés para hablantes nativos) y textos simplificados (adaptaciones de textos auténticos para estudiantes de inglés como segunda lengua). Los autores utilizan Coh-Metrix [Graesser et. al, 2011] para extraer métricas para hacer su estudio.

[4] www.weeklyreader.com

El problema que se trata de solucionar es el poco análisis experimental que se tiene sobre el uso de textos auténticos o simplificados en la enseñanza de inglés como segundo idioma para estudiantes de nivel básico o intermedio. En este caso, los expertos solo hacían recomendaciones en base a métricas convencionales como fórmulas o conteo de palabras clave. Mediante el análisis se explica la validez de estas investigaciones.

Se usó Coh-Metrix para la extracción de medidas de cohesión causal, conectores y operadores lógicos, correferencia, densidad de partes principales del texto, polisemia e hiperónimos, complejidad sintáctica, y también las métricas básicas de frecuencias. En base a estos valores, el autor pudo analizar textos de alumnos de los cursos de inglés como segunda lengua y contrastar las diferencias en complejidad entre los textos auténticos y simplificados.

El estudio concluye con resultados experimentales, en las cuales el autor extrajo métricas de los textos de forma automática por medio del uso de la herramienta Coh-Metrix y por medio de métodos estadísticos logró demostrar la mayor correferencia y cohesión de los textos simplificados, así como su poca diversidad de vocabulario y menos causalidad, y menor dependencia en operadores lógicos. Además de ello, se demuestra que no existe diferencia entre el nivel de abstracción y ambigüedad de los textos auténticos y simplificados [Cossley et al. 2007].

4.2 Productos comerciales para resolver el problema

Se encontraron los siguientes productos comerciales:

- **Lexile**

  Para el análisis de textos se tiene una escala de lectura de **MetaMetrix**, llamada **Lexile**, la cual desarrolló un ranking entre la complejidad de textos en inglés. Con esta misma escala se mide la capacidad de comprensión lectora de los evaluados y se da un rango en el cual este evaluado puede tener compresión del texto. Esta solución solo está disponible al 100% en inglés.

  Al ser una herramienta comercial, no se conocen las métricas utilizadas, solo se menciona el uso de conteo de palabras, vocabulario y propiedades sintácticas utilizadas [METAMETRIX, 2013].

- **Microsoft Word**

  En versiones antiguas de este programa (hasta la versión 2000) contaba con indicadores para el idioma español. En versiones recientes, cuenta tan sólo con métricas descriptivas como contadores de palabras, caracteres, párrafos, entre otros que no brindan mayor información para el análisis del texto.

4.3 Productos no comerciales (de investigación) para resolver el problema

Se mencionarán las herramientas para el idioma español. Tener en cuenta que son modelos aproximados:

- **Inflesz**

  Es una aplicación libre diseñada en lenguaje C++ para el análisis de fragmentos o textos completos en español. Posee un número limitado de métricas comparadas con las herramientas mencionadas anteriormente, pero es la herramienta que presenta métricas más útiles para el idioma español [Inflesz, 2007]:

  - Palabras
  - Sílabas
  - Frases (oraciones)
  - Promedio de sílabas por palabra
  - Promedio de palabras por frase
  - Métrica Flesch-Szigrizt
  - Grado en la escala Inflesz
  - Correlación Word
  - Fórmula Flesch-Fernandez Huerta

Para la extracción de métricas se tienen métricas desarrolladas en inglés y adaptaciones de estas al portugués. Se mencionan las más importantes:

- **Coh-Metrix**

  Es una herramienta disponible en la web que analiza textos en inglés usando métricas lingüísticas. Fue desarrollada por investigadores de la Universidad de Memphis y ya se encuentra en su versión 3.0. Esta cuenta con 110 métricas en su versión gratuita, 40 más que su segunda versión, y con más de 200 métricas (incluye los anteriores) para el uso de investigadores. Coh-Metrix calcula métricas que evalúan la cohesión, coherencia y dificultad de comprensión de un texto [Coh-Metrix, 2011]. Dado que el Coh-Metrix posee muchas métricas, estos se encuentran clasificados en 11 grupos, los que a continuación se irán describiendo:

  - ***Descriptives*:** Sirven para poder corroborar que la información obtenida en el análisis tiene sentido y para poder observar patrones de datos. Gran parte de estos son contadores (palabras, oraciones, párrafos, entre otros)
  - ***Text Easability Principal Component Scores*:** Estos dan una imagen más precisa de la facilidad o dificultad que se observan en las características lingüísticas del texto.
  - ***Referential Cohesion*:** Estas métricas miden la capacidad que podría tener un lector para conectar las ideas de un texto entre proposiciones y oraciones del texto.
  - ***Latent Semantic Analysis*:** Miden las coincidencias o parecidos entre oraciones y párrafos.
  - ***Lexical Diversity*:** Miden la relación entre el número de tipo de palabras y el número de palabras totales para mostrar si el texto tiene una alta cohesión.
  - ***Connectives*:** Mide la incidencia de los conectores en el texto en las cinco clasificaciones que le da a estos (causales, lógicos, adversativos, temporales y aditivos), además, tiene una clasificación de conectores positivos y negativos.
  - ***Situation Model*:** Tienen que ver con la representación mental que el lector puede hacer sobre el texto.

- ***Syntactic Complexity*:** Analizan la sintaxis de las oraciones para poder ver si estas poseen mucha carga de palabras, lo que las hace complejas.
- ***Syntactic Pattern Density*:** Calcula la incidencia de los diferentes tipos de patrones que existen en el texto (frases nominales, frases verbales, frases adverbiales, entre otros).
- ***Word Information*:** Muestra los puntajes de incidencia de los diferentes tipos de palabras, analizando primero de qué tipo es cada palabra en dicho contexto.
- ***Readability*:** Mide la legibilidad del texto, esencialmente utilizando la métrica conocida como Flesch Reading Ease (FRE), una de las métricas más usadas para analizar esta característica.

Este software utilizó diversas herramientas para poder calcular los más de 100 métricas que posee, entre ellos una base de datos de más de 17 millones de palabras (y sus frecuencias), una lista de conectores clasificados, entre otros. Una de las aplicaciones de Coh-Metrix es el análisis de documentos clínicos para que los doctores puedan emitir reportes clínicos que los pacientes puedan entender con facilidad [Graesser, 2005].

- **Coh-Metrix-Port**

Este software está basado en el Coh-Metrix para el idioma inglés en su versión 2.0, pero no en su totalidad ya que solo implementa 40 métricas para el análisis de textos. Es una herramienta web y de código fuente libre. Fue desarrollado en Brasil por el Núcleo Interinstitucional de Lingüística Computacional (NILC) para el análisis de textos en portugués, pero este pertenece realmente a un proyecto mucho más grande para la simplificación de textos complejos. Las métricas que se adaptaron e implementaron para el portugués son clasificadas en 7 bloques:

- **Contadores básicos:** Datos básicos del texto como el número de palabras, oraciones, y el promedio de estos como número de palabras promedio por oración.
- **Constituyentes:** Analiza la incidencia de sintagmas nominales, el número de modificadores por sintagma nominal y el número de palabras antes de verbos principales.
- **Frecuencias:** Muestra las frecuencias de los diferentes tipo de palabra que existen en el texto.
- **Conectores:** Mide la incidencia de los diferentes tipos de conectores. Tiene la misma clasificación del Coh-Metrix inglés (conectores aditivos, causales, temporales, lógicos y aditivos).
- **Operadores Lógicos:** Cantidad de los diferentes operadores lógicos en el texto como y, o, si y variaciones de la negación.
- **Pronombres:** Cantidad de pronombres promedio de sintagmas en el texto y su incidencia.
- **Ambigüedades:** Mediante un diccionario reducido de palabras (20 mil) puede medir el grado de ambigüedad no de las palabras, sino de los tipos de palabras (verbos, adjetivos, adverbios, sustantivos).

Como se mencionó anteriormente, una de las principales aplicaciones del Coh-Metrix-Port fue la de ayudar al análisis de complejidad de textos y la simplificación de los mismos [Scarton, 2010].

- **QUAID**

  Esta es una aplicación que permite el análisis de preguntas en el idioma inglés. Su aplicación básica es la de poder analizar cuestionarios y notar si podrían existir problemas de comprensión por parte del lector. Con QUAID se encontraron principalmente cinco problemas para la comprensión de las preguntas:

  - Términos muy técnicos
  - Predicados imprecisos o términos relacionados
  - Frases nominales ambiguas
  - Sintaxis compleja
  - *Working memory load*

  La última de estas se refiere a la carga mental que debería tener el lector para poder entender todas las relaciones establecidas en la pregunta. Por ejemplo, si se dan muchas clasificaciones para escoger y, posiblemente, cruces entre ellas, el lector tendría que crear una matriz mental de lo que implicaría cada cruce de estas clasificaciones. Esto puede complicar el entendimiento de la pregunta y, a su vez, dificultar una respuesta o quitarle la veracidad a esta misma [Graesser, 2005].

- ***Linguistic Inquiry And Word Count***

  Herramienta de análisis de textos en inglés. Se creó con el fin de analizar textos narrativos escritos por víctimas de eventos traumáticos. Permite predecir cuán bien el paciente puede convivir o superar el trauma, además de aproximar el número de visitas necesarias para que este mejore [Graesser, 2005].

A continuación se presenta un resumen de las herramientas e investigaciones revisadas en el estado del arte. Se tiene en cuenta si es de licencia comercial, los idiomas en los que están disponibles, técnicas de clasificación de complejidad que utilizan y las métricas en que basan su clasificación:

Tabla 2 Comparación de herramientas y análisis. Fuente: propia

Característica/Producto	Libre	Idioma	Técnica de clasificación	Métricas
Tanaka	Sí	Inglés, Japonés	Ordenamiento y SVM	Uso de vocabulario
Schwarm, Petersen	Sí	Inglés	SVM y Regresión,	Conteo de propiedades de árboles de análisis sintáctico.
Crossley	Sí	Inglés	Manual	Coh-Metrix
Coh-Metrix	Sí	Inglés	Solo brinda métricas.	Contadores básicos, Constituyentes, Frecuencias, Conectores, Operadores Lógicos, Pronombres, Ambigüedades

Inflesz	Sí	Español	Solo brinda métricas.	Palabras, Sílabas, Frases (oraciones), Promedio de sílabas por palabra, Promedio de palabras por frase, Métrica Flesch-Szigrizt, Grado en la escala Inflesz, Correlación Word
Word	No	Inglés, Español(200)	Solo brinda métricas.	Contadores de palabras, caracteres, párrafos
Coh-Metrix-Port	Sí	Portugués Brasil.	Solo brinda métricas.	Contadores básicos, Constituyentes, Frecuencias, Conectores, Operadores Lógicos, Pronombres, Ambigüedades
Graesser	Sí	Inglés	Solo brinda métricas.	Conteo de palabras y vocabulario.
Lexile	No	Inglés	No especifica.	No especifica.
Quaid	Sí	Inglés	No especifica.	Términos muy técnicos Predicados imprecisos o términos relacionados Frases nominales ambiguas Sintaxis compleja Working memory load

4.4 Conclusiones sobre el estado del arte

De las investigaciones y herramientas presentadas se puede observar que gran parte del trabajo hecho en el área de procesamiento de texto para la clasificación por complejidad está hecho para tratar textos en inglés. También que la automatización de la extracción de métricas permite realizar análisis experimentales que ayuden a confirmar hipótesis de investigadores de lingüística. De esto se identifica que existen recursos base para el desarrollo de estas investigaciones para cualquier idioma, lo que nos da la oportunidad de utilizar estos recursos para el desarrollo en español.

Es así que, por medio de este trabajo de fin de carrera, se desea hacer una herramienta de análisis de complejidad de textos que permita el análisis experimental por medio de la extracción automática de métricas del texto **en idioma español**, y con esto brindar una alternativa de clasificación en textos de fácil y difícil comprensión que se pueda realizar de forma numérica y automática.

CAPÍTULO 2

1 Objetivo general

- Implementar una herramienta de análisis y clasificación de complejidad de textos en español.

2 Objetivos específicos

1. Adaptar las métricas psicolingüísticas utilizadas en el idioma inglés para el idioma español.
2. Implementar una herramienta de extracción de métricas psicolingüísticas de legibilidad de un texto en español.
3. Recopilar un corpus[5] de textos en español clasificados manualmente según una escala de complejidad.
4. Implementar un clasificador de complejidad de textos en español que utilice métricas psicolingüísticas de legibilidad y métodos de aprendizaje de máquina.
5. Implementar una herramienta informática de clasificación automática de complejidad de textos en español.

3 Resultados esperados

1. Para el objetivo 1: Lista de 108 métricas psicolingüísticas del Coh-Metrix (inglés) analizadas y priorizadas para su implementación según recursos necesarios y disponibles.
2. Para el objetivo 2: Herramienta de cálculo de métricas psicolingüísticas de legibilidad de un texto en español con al menos 48 métricas implementadas.
3. Para el objetivo 3: Corpus de 100 textos (cuentos) en español anotados manualmente y clasificados según su complejidad (simple y complejo).
4. Para el objetivo 4: Modelo de clasificación automática de complejidad de textos en español basado en experimentación numérica con una precisión mayor a 60%.
5. Para el objetivo 5: Servicio web de clasificación automática de textos en español que utilice el mejor modelo de clasificación obtenido luego de la experimentación numérica.

[5] Corpus: Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios que pueden servir de base a una investigación.[RAE, 2010]

4 Herramientas, métodos y procedimientos

4.1 Mapeo

A manera de resumen, se listarán las metodologías, métodos o procedimientos que se utilizaron para cada uno de los resultados esperados para luego explicar con más detalle cada uno.

Resultado Esperado 1	Metodología, método o procedimiento
Lista de 108 métricas psicolingüísticas del Coh-Metrix (Inglés) analizadas y priorizadas para su implementación según recursos necesarios y disponibles.	Se revisó la documentación del Coh-Metrix 3.0 y Coh-Metrix-Port y analizó si cada una de las métricas podía ser adaptada al idioma español en base a similitudes morfológicas, sintácticas, semánticas y gramaticales de los lenguajes. Se adaptó, de forma manual, al idioma español la lista de métricas filtradas luego de su análisis.

Resultado Esperado 2	Metodología, método o procedimiento
Herramienta de cálculo de métricas psicolingüísticas de legibilidad de un texto en español con al menos 48 métricas implementadas.	Se implementó la extracción automática de las medidas psicolingüísticas de legibilidad adaptadas para el español. Para ello, se utilizó Freeling, que es un paquete de software que provee herramientas de PLN[6] para diferentes idiomas incluyendo el español.

Resultado Esperado 3	Metodología, método o procedimiento
Corpus de 100 textos (cuentos) en español anotados manualmente y clasificados según su complejidad (simple y complejo).	Se buscó un corpus de textos en español de género literario y se anotó según su complejidad.

Resultado Esperado 4	Metodología, método o procedimiento
Modelo de clasificación automática de complejidad de textos en español basado en experimentación numérica con una precisión mayor a 60%.	Se utilizó el paquete WEKA para realizar una experimentación usando los diferentes algoritmos de aprendizaje de máquina de tipo supervisado que este posee y así poder elegir el mejor modelo de clasificación. Esta comparación fue realizada en base a estadísticas de precisión.

[6] PLN: Procesamiento de lenguaje natural (o NLP, por sus siglas en inglés)

Resultado Esperado 5	Metodología, método o procedimiento
Servicio web de clasificación automática de textos en español que utilice el mejor modelo de clasificación obtenido luego de la experimentación numérica.	Se integraron todas las herramientas desarrolladas y se desarrolló una interfaz web que contiene la aplicación integrada.

4.2 Coh-Metrix

Para saber qué métricas psicolingüísticas se iban a implementar para la herramienta en español se tomaron como base las aplicaciones Coh-Metrix 3.0 (para idioma inglés) y Coh-Metrix Port (para idioma portugués). La primera de estas fue tomada como referencia por poseer más de 100 métricas debidamente documentadas para análisis de textos en inglés. De manera similar, la documentación de la segunda aplicación fue revisada para poder encontrar cuáles fueron las dificultades o ventajas que tuvieron al realizar las adaptaciones del inglés al portugués de cada una de las métricas que se implementaron (40 métricas adaptadas del Coh-Metrix 2.0).

En dicho proceso de revisión de la documentación de cada una de las aplicaciones, se revisaron las características necesarias para poder calcular cada métrica y encontrar las similitudes entre los lenguajes tanto sintácticas, semánticas y gramaticales. En base a dicha revisión se analizó si era posible adaptar o no cada métrica al idioma español.

Esta tarea se realizó de esta forma porque se sabe que Coh-Metrix 3.0 tiene una amplia cantidad de métricas implementadas y, además, que es usada también por investigadores del área [Coh-Metrix-3, 2011]. También, Coh-Metrix-Port fue de gran ayuda dado que este se encuentra basado en la herramienta mencionada anteriormente y aportó las diversas formas que se tuvieron para adaptar diferentes métricas al portugués y tomarlas como ejemplo para la adaptación al idioma español en los casos que era posible [Scarton, 2010].

4.3 Freeling

Para el desarrollo de la herramienta que calcule las métricas adaptadas en el paso anterior se necesitaron de diversos recursos los cuales fueron analizados en esta etapa. El principal recurso que se utilizó fue un *parser* para español para poder obtener diversas características de las palabras o frases del texto el cual es brindado por Freeling porque tiene implementadas la mayoría de funcionalidades necesarias para español.

Freeling es una librería que provee servicios de análisis de lenguaje [Freeling, 2012]. Dentro de estos servicios provee identificación de lenguaje, separación de oraciones, análisis morfológico, marcado de texto, desambiguación y resolución de correferencia.

4.4 Origen del corpus de texto

Se buscó un corpus de textos en español que sea adecuado para el proyecto. Se inició primero una búsqueda de un corpus en español que sea estándar para investigaciones y posea las características lo más parecida posibles a las métricas que se implementaron, además de la clasificación que tienen. Luego se recopilaron textos de diversas fuentes online que contenían cuentos y fábulas.

4.5 WEKA

Una vez implementadas las métricas psicolingüísticas y obtenido el corpus de textos periodísticos en español clasificados se realizó una serie de experimentos para obtener el mejor modelo de clasificación. Para esto, se utilizó el paquete WEKA, que contiene varias rutinas de aprendizaje de máquina y es desarrollado por la Universidad de Waikato, por medio de su meta clasificador y filtros ya implementados. [WEKA, 2013]

Se buscarán qué modelos son usados generalmente para esta tarea (clasificación de textos) y se hará una experimentación en base a un mismo corpus de textos para ver cuál de dichos modelos es el que aprendió mejor a clasificarlos comparando la clasificación real con la arrojada por el modelo de aprendizaje. Al finalizar dicho experimento se elegirá al mejor de todos en base a los resultados y se tomará este para la herramienta de clasificación. Estos modelos serán generados por algoritmos de aprendizaje de máquina de la herramienta WEKA.

A pesar que se encuentre un modelo que siempre devuelva muy buenos resultados según varias investigaciones, se debe hacer una experimentación porque podría ser que este no lo sea para las características implementadas sino algún otro modelo [Bishop, 2006]. Algunos de los modelos que formarán parte del experimento son los siguientes:

- Máquina de vector de apoyo (SVM)
- Gaussiana de regresión
- Árbol de decisión de aprendizaje

4.6 Construcción del servicio web

Para esta etapa ya se tenían desarrolladas ambas herramientas: extractor de métricas psicolingüísticas de legibilidad de textos en español y el clasificador de textos; sin embargo, estas trabajaban por separado y no se encontraban integradas por lo que aquí se procedió a su integración. Además, se diseñó la interfaz web e implementó el servicio web de la herramienta final.

Para poder ofrecer una mejor herramienta al usuario y no dos por separado, se procedió a la integración de estas ayudando así tanto a redactores como a lectores saber con anticipación si un texto podrá ser leído con facilidad o no. Además, al no existir una herramienta que brinde dichas funcionalidades para el idioma español, se tendrá un gran aporte al área y será de fácil acceso al tener un portal web para su uso.

4.7 Metodología para el desarrollo de la herramienta.

Para el desarrollo de la herramienta se utilizó la metodología de cascada incremental iterativa. El proyecto es simple de manera visual ya que solo requiere de la entrada de un texto y la confirmación del envío. Se desarrollaron las etapas de especificación de requisitos de las funcionalidades del clasificador. Diseño e implementación para el extractor de métricas y la interfaz para el uso de los algoritmos de aprendizaje de máquina y pruebas.

Se elige esta metodología por la simplicidad en los requerimientos y la interacción con este. Con esta metodología propone enfocarse más en la implementación de la extracción de las métricas e interacción con los algoritmos de aprendizaje de máquina ya que estos factores, en conjunto con la recopilación del corpus, son los de mayor complejidad en el proyecto [Weitzenfeld, 2004].

5 Alcance

El proyecto es de Ciencias de Computación, de la rama de Procesamiento de Lenguaje Natural. Es de tipo investigación aplicada, ya que se plantea dar una alternativa al problema de clasificación automática de complejidad de textos en español mediante métodos de aprendizaje estadístico con ayuda de métricas psicolingüísticas de legibilidad. No se pretende dar una solución exacta a este problema. Esto se debe a que es un problema abierto de forma exacta y los métodos estadísticos ofrecen una aproximación práctica [Manning, 1999].

Se toma a los textos literarios (cuentos) en español como corpus base de estudio para el análisis. Se elige este tipo de textos porque ya existe una división natural según el público al que está dirigido, desde fábulas para niños hasta relatos para adultos, pudiendo tomarse esto como complejidad simple y compleja, respectivamente. También se toma al idioma español, por la escasez de herramientas de clasificación de complejidad textual en este idioma, el conocimiento que se tiene sobre este y la disponibilidad de expertos en lingüística en el área dentro de la universidad donde se desarrollará el proyecto.

En cuanto a las escalas de complejidad a utilizarse, se plantean 2 clases; *simple*, texto poco elaborado, sin uso de conectores lógicos avanzados y de vocabulario corto; y *complejo*, textos con un uso de vocabulario más extenso y de construcciones elaboradas.
Con el desarrollo de esta herramienta basada en métodos de aprendizaje estadístico, se busca acercarse a un modelo para la complejidad textual en español más detallado en cuanto a clases de complejidad y tomando inicialmente métricas psicolingüísticas de legibilidad.

5.1 Limitaciones

Los métodos de clasificación a utilizar serán de tipo supervisado, por lo que son totalmente dependientes del corpus de entrenamiento y de la definición inicial de las clases. No se plantea el uso de métodos que permitan adaptabilidad ya que esto excede al alcance del proyecto.

Se utilizaran 2 clases de complejidad por la dificultad en la anotación del corpus, ya que el corpus inicial de entrenamiento necesita ser extraído. Esto lleva consigo un margen de subjetividad en la evaluación y al aumentar el número de clases la subjetividad también aumenta.

Como en las investigaciones revisadas, se necesitará un corpus de estudio en español que del cual se extraigan las métricas y sobre el cual se pueda realizar el entrenamiento y evaluación del clasificador. Se puede observar que en las investigaciones revisadas este corpus es de origen periodístico, dada la diversidad y facilidad de acceso que se tiene a este recurso, sin embargo, para el proyecto se tomará el género literario dado que no existe una fuente confiable con textos periodísticos que puedan ser catalogados como simples.

Se utilizará un subconjunto de métricas psicolingüísticas de Coh-Metrix que tenga equivalente en español, tratando de implementar el total de éstas. Esto es porque estas métricas fueron diseñadas para inglés y algunas estructuras gramaticales no tienen una equivalencia directa en español.

5.2 Riesgos

Tabla 3 Riesgos identificados del proyecto

Riesgo identificado	Impacto para el proyecto	Medidas Correctivas para mitigar
Alta complejidad en la adaptación de métricas.	Retraso en la etapa de desarrollo de métricas.	Adaptar e implementar las métricas simples inicialmente, para dejar tiempo a las más complejas
Retraso en la clasificación manual de los textos para entrenamiento.	Retraso en las pruebas generales del algoritmo de clasificación.	Priorizar la recolección y clasificación manual del texto desde el inicio del proyecto.
Conocimiento insuficiente en lingüística.	Retraso general en el desarrollo de las métricas afectadas.	Contar al menos con 1 experto en lingüística de apoyo en el proyecto.
No contar con todos los recursos necesarios para la extracción de métricas en español.	Reducción en el número de métricas implementadas y posible pérdida de precisión del clasificador.	Monitorear el desarrollo de las métricas más complejas e identificar las herramientas necesarias.

6 Justificación y viabilidad

En la siguiente sección se verificará la viabilidad y se mencionará la justificación para llevar a cabo el proyecto.

6.1 Justificación del proyecto de tesis

La lectura es parte fundamental para la educación, ya que se está aprendiendo a leer y leyendo para aprender. Es por ello que se necesitan herramientas para dar un adecuado orden de dificultad para lograr un escalamiento de complejidad de los textos estudiados y lograr exposición progresiva a textos complejos [Fisher et. al, 2012].

Este proyecto beneficia a los docentes y alumnos, ya que brinda acceso a herramientas automatizadas de análisis de complejidad de textos en español. Dentro de este análisis se abarcan más que métricas de frecuencias de palabras. Se busca extraer medidas más complejas en cuanto a cohesión. Con esta accesibilidad de herramientas se busca que los alumnos y docentes de idioma español tengan igual facilidad en cuanto a escalamiento y análisis de complejidad de textos como los ya aplicados para el idioma inglés.

Adicionalmente, por la metodología utilizada, se puede extender el proyecto a más escalas de complejidad de textos basadas en métricas de cohesión. También, se puede llegar hacer un análisis de la estructura del corpus en base a métricas de cohesión por medio de la aplicación de algoritmos de aprendizaje no supervisado. Finalmente, el motor de extracción de métricas de cohesión y estadísticos podría ser utilizado para distintos análisis en el área de procesamiento de lenguaje natural en español.

6.2 Análisis de viabilidad del proyecto de tesis

El tiempo estimado para el proyecto es de 4 meses, en el que se usarán herramientas gratuitas lo que evitará que se tengan limitaciones financieras y de accesibilidad en el proyecto. Además, se estima que se podrá desarrollar el proyecto en el tiempo propuesto con los conocimientos del área de procesamiento de lenguaje natural que se mencionaron para cada una de las etapas previamente definidas. Se tiene en cuenta también que el proyecto estará a cargo de dos personas y que se podrá distribuir tareas en paralelo o subdividir algunas, de esta forma no habrá retrasos ni largos periodos en los que el proyecto esté en progreso.

CAPÍTULO 3

1 Introducción

En este capítulo se hablará de la herramienta de extracción de métricas (o Coh-Metrix-Esp) y los pasos que se siguieron para su implementación. Se iniciará explicando el proceso de filtrado de métricas a implementar para luego explicar cada uno de los módulos que contendrá dicha herramienta, así como la descripción de cada una de las métricas mencionando los recursos necesarios para cada una que fueron utilizados.

2 Lista de métricas psicolingüísticas de legibilidad

Para empezar con el proyecto se necesitaban principalmente las métricas que serían implementadas, pero dado que no se tenían conocimientos suficientes para crear métricas propias se optó por tomar como referencia las métricas que brinda Coh-Metrix 3.0 (Inglés). Esta herramienta cuenta con 108 métricas divididas en 11 grupos.

En primer lugar, se analizó cada uno de las métricas que brindaba esta herramienta y se procedió a ver si existían equivalentes para el idioma español o necesitaban ser adaptadas para este.

En segundo lugar, para cada una de las métricas se verificó usando Coh-Metrix-Port si se había implementado en dicha herramienta y si había sido necesaria su adaptación al idioma portugués, en cuyo caso se observaría si hubo complicaciones en dicho proceso. Este análisis puede ser verificado en la cuarta columna de la Tabla 4[7]. Con esto se lograron filtrar muchas métricas que no serían implementadas para el proyecto.

Adicionalmente, se tienen las columnas *Recursos necesarios* y *Observaciones* donde se colocan las herramientas necesarias y algunas posibles complicaciones para implementar dicha métrica, respectivamente.

Con las métricas que no habían sido implementadas por el Coh-Metrix-Port se realizó una segunda revisión en caso existan algunas que puedan ser fácilmente implementadas en el idioma español. Esto se realizó dado que Coh-Metrix-Port está basado en una versión antigua de Coh-Metrix (Inglés) la cual contenía menos de cien métricas.

Tabla 4 Extracto de análisis de métricas

Grupo	Métrica	Situación	En Coh-Metrix-Port	Recursos necesarios	Observaciones
Descriptives	Número de párrafos	Se implementará	Sí	Ninguno	
	Número de oraciones	Se implementará	Sí	Freeling	
	Número de palabras	Se implementará	Sí	Freeling	
	Número de oraciones por párrafo (media)	Se implementará	Sí	Freeling	
	Número de oraciones por párrafo (desviación	Se implementará	No	Freeling	

[7] La tabla completa se encuentra en el Anexo 1

	estándar)				
	Número de palabras por oración (media)	Se implementará	Sí	Freeling	
	Número de palabras por oración (desviación estándar)	Se implementará	No	Freeling	
	Número de sílabas por palabra (media)	Se implementará	Sí	Freeling	
	Número de sílabas por palabra (desviación estándar)	Se implementará	No	Freeling	
	Número de letras por palabra (media)	Se implementará	No	Freeling	
	Número de letras por palabra (desviación estándar)	Se implementará	No	Freeling	

Finalmente, luego de realizar los filtros respectivos usando las herramientas previamente mencionadas, y además verificar que existan recursos necesarios para la implementación de las métricas, se procedió a llenar la columna *Situación* la cual solo estaría llena si dicha métrica se va implementar para el presente proyecto.

En la Tabla 5 se puede observar de forma general el número de métricas implementadas por cada grupo.

Tabla 5 Métricas a implementar por grupo

Grupo	Seleccionadas	Total métricas
Descriptives	11	11
Text Easability	0	16
Referential Cohesion	14	14
LSA	0	8
Lexical Diversity	2	4
Connectives	6	9
Situation Model	0	8
Syntactic Complexity	1	7
Syntactic Pattern Density	3	8
Word Information	10	22
Readability	1	3
Total	**48**	**108**

Como se puede observar, son implementadas un total de 48 métricas de las 108 que posee Coh-Metrix 3.0.

3 Herramienta de cálculo de métricas psicolingüísticas

En este capítulo se describe el resultado 2 que cubre al objetivo 2: implementar una herramienta de extracción de métricas psicolingüísticas de un texto en español. Para ello fue necesario implementar las 48 métricas previamente seleccionadas y analizadas en el Resultado 1. Se presenta el análisis clasificado en base a las secciones de Coh-Metrix 3.0.

1. Freeling y diseño base:

Por el análisis previo realizado en Resultado 1[8], se notó que para 24 de 48 métricas se necesitaría de un marcador morfológico. Es por ello que se elige a Freeling 3.1 [Freeling, 2012], por las funcionalidades que ofrece para el idioma español (Tabla 6) y su facilidad de integración como librería base para la aplicación.

Tabla 6 Funcionalidades de Freeling

Partes Utilizadas
Tokenization
Sentence Splitting
Number Detection
Date Detection
Morphological dictionary
Multiword Detection
Part of Speech tagging
Shallow parsing

Freeling es publicado bajo la licencia GNU GPL, por lo que tanto el código fuente y los ejecutables están disponibles en su página web [Freeling, 2012].

Esta herramienta nos brinda una API en Java y estructuras básicas para el marcado y análisis de lenguaje natural. Después de evaluar la API y diseñar alternativas para el marco de trabajo para el desarrollo de la herramienta de cálculo de métricas, se identificó la necesidad de estructuras de más alto nivel, así como el desacoplamiento de Freeling y su configuración del cálculo de las métricas. Es por ello que se diseñó la arquitectura que se presenta en la Figura 6.

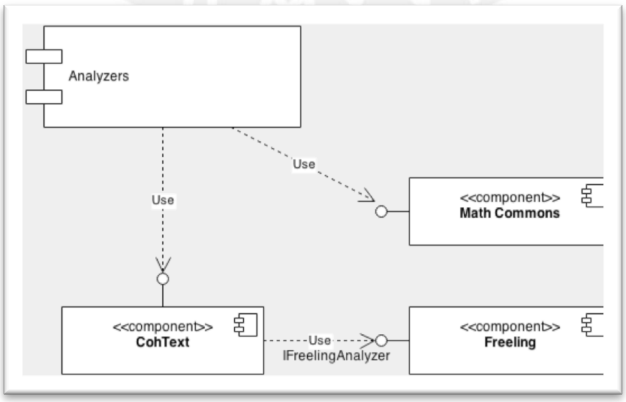


Figura 6 Diagrama de despliegue

[8] El resultado 1 es la lista completa de las métricas de Coh-Metrix (Inglés) analizada para su implementación. La lista se encuentra en Anexo 1.

- Se implementaron las clases CohText, CohParagraph, Sentence y Word. Estas contienen los métodos para delegar los cálculos a Freeling y utilidades para la interacción entre objetos.
- Se accede a Freeling bajo la interfaz IFreelingAnalyzer, que maneja lo necesario para su configuración.
- Cada Analyzer cumple con el desarrollo de las métricas seleccionadas de Coh-Metrix y solo depende de los pre-cálculos hechos en CohText. Estos Analyzer realizan el cálculo de métricas y las devuelven en conjunto con el código de la métrica de Coh-Metrix (Ejm, DESPC, Número de párrafos).
- Para el cálculo de las medidas estadísticas se utiliza la librería MathCommons.

2. Módulo de Métricas Descriptivas

Se decide implementar todas las métricas de este módulo. En esta sección se explicarán los detalles de implementación y dificultades de las siguientes métricas:

- **Número de Párrafos (DESPC):** Para esta métrica se tomó a un párrafo como el texto entre 2 cambios de línea o el inicio y fin del texto. La separación de párrafos dentro del texto se hace dentro de la creación de CohText, donde se crea una lista y se insertan los párrafos en forma de clase CohParagraph, el cálculo de la métrica retorna la longitud de esta lista.

- **Número de Oraciones (DESSC):** Se utiliza el módulo de separación (splitter) de Freeling. La separación y creación de oraciones es hecha por Freeling en la llamada inicial hecha dentro de CohText. Se tomaron como fines de oración los cambios de línea, signos de puntuación y se permite la separación de oraciones dentro de anidaciones de paréntesis.

- **Número de Palabras (DESWC):** Se utiliza el módulo de análisis morfológico de Freeling. En nuestra implementación se excluyen como palabras a los números y los signos de puntuación.

- **Media y desviación estándar de la longitud de los párrafos (DESPL, DESPLd):** Se utiliza el splitter de Freeling. Se hace el cálculo en base al número de oraciones que se tiene en cada párrafo.

- **Media y desviación estándar del número de palabras por oración(DESSL, DESSLd):** Se utiliza la misma clasificación de palabra que en (DESWC) y se cuenta cada una de estas por oración detectada.

- **Media y desviación estándar de la media de sílabas por palabra (DESWLsy, DESWLsyd):** Para su implementación se desarrolló un separador en sílabas basado en expresiones regulares [Warck, 2005]. Se agruparon las reglas tomadas de la RAE y se implementa esto en la clase SylableSplitter.

- **Media y desviación estándar de la media de letras en cada palabra (DESWLlt, DWSLltd):** Como en (DESWC) se toma a cada palabra y se calcula el número de letras en base a los índices de inicio y fin brindados por Freeling.

3. Módulo de Cohesión Referencial

Se decidió implementar la totalidad de métricas de este módulo. Estas métricas miden las conexiones que existen en el texto, cada una de las métricas evalúa un tipo de conexión entre pares adyacentes (sufijo '1' en el nombre de la métrica) y entre todos los pares de oraciones (sufijo 'a' en el nombre de la métrica). Es por ello que todas comparten el mismo iterador y se independizan las funciones que evalúan cada uno de los tipos de cohesión. Se utiliza el módulo de etiquetado gramatical de Freeling para obtener las etiquetas y formas básicas de las palabras. Tomando esto en cuenta, las métricas implementadas son:

- **Superposición de sustantivos (CRFNO1, CRFNOa):** Se comparan los conjuntos de palabras de 2 oraciones y se retorna que tienen una superposición siempre y cuando se tenga la misma forma de sustantivo en las oraciones evaluadas.

- **Superposición de Argumentos (CRFAO1, CRFAOa):** Se comparan los conjuntos de palabras de las 2 oraciones evaluadas y se detecta una superposición de argumentos en caso que compartan algún sustantivo o pronombre.

- **Superposición de Temas (Stems) (CRFSO1, CRFSOa):** Se relaja la restricción de CRFAO1 y CRFAOa, se detecta una superposición de temas cuando el lema de alguna palabra de contenido (sustantivos, verbos, adjetivos, adverbios) es compartida por las 2 oraciones evaluadas.

- **Superposicion de palabras de contenido (CRFCWO1, CRFCWO1d, CRFCWOa, CRFCWOad):** Se relaja la restricción de Temas y se comparan todas las palabras de contenido entre 2 oraciones, detectando una superposición en caso compartan la misma palabra. Adicionalmente, se mide la desviación estándar del caso de oraciones adyacentes (CRFCWO1d) y al comparar todas los pares de oraciones (CRFCWOad).

- **Superposición de anáforas (CRFANP1, CRFANPa):** Se detecta una anáfora si se encuentra un pronombre de la primera oración que se refiere a un sustantivo o pronombre de la segunda. Esto se hace comparando los géneros y tiempo de las palabras; ambas etiquetas son calculadas con el POS tagger de Freeling.

Dentro de este módulo se observó que la comparación de todos los pares de oraciones en algunos casos no es de relevancia, ya que algunas condiciones relajadas no brindan información adicional porque a medida que las oraciones se separan la relación entre estas también se debilita. Se puede llegar a analizar la relevancia de esta localidad en trabajos posteriores.

4. Módulo de Diversidad Léxica

En este módulo se calculan métricas referentes a la cantidad de palabras diferentes dentro del texto. Se implementaron 2 de las 4 métricas originales.

- **Porcentaje de palabras de contenido diferentes (LDTTRc):** Se contó las palabras de contenido diferentes (se excluyen signos de admiración, números y fechas) y esta cantidad fue dividida por el total de palabras de contenido.

- **Porcentaje de palabras diferentes (LDTTRa):** Se hace el mismo cálculo de LDTTRc, pero sin la restricción de palabras de contenido.

5. Módulo de Conectores

Las métricas incluidas en este módulo están relacionadas directamente a los conectores y sus diferentes clasificaciones. Para poder implementar estas métricas se tuvo que crear una lista que contenga los conectores con sus respectivas clasificaciones porque la herramienta principal usada, Freeling, no brindaba la identificación de estos al ser muchos de ellos compuestos por más de una palabra. En primer lugar, se procedió a revisar fuentes en internet que brindaran listas de conectores y su clasificación para poder extraer la mayor lista de conectores posibles y luego se realizó una comparación entre estas para poder identificar cuáles eran las clasificaciones posibles de cada uno, dado que algunos de estos podrían estar incluidos en más de un grupo de conectores. Además, se comparó esta con la clasificación de conectores que brindaba el libro Nueva Gramática[9] elaborado por la RAE (Real Academia Española).

Una vez elaborada la lista[10] y las posibles clasificaciones para cada elemento en dicha lista, se procedió a elaborar un archivo de texto que contenía dos líneas por cada conector. La primera línea contenía al conector y la siguiente correspondía a una cadena de caracteres con iniciales mayúsculas según las clasificaciones a la que pertenecía. Por ejemplo, para el conector *entonces* se tenía la cadena *CT* como sus clasificaciones, es decir que este podía ser un conector causal o temporal. Dicho archivo era cargado a la memoria de la aplicación para poder calcular las métricas involucradas en este módulo, en total 6 de las 9.

Por último, se tiene el término *incidencia* para el cálculo de estas métricas que consiste en el número de unidades encontradas por mil palabras.

- **Incidencia de todos los conectores (CNCAll):** Se realiza un conteo de todos los conectores en el texto que se encuentren en el archivo de conectores elaborado.

- **Incidencia de conectores causales (CNCCaus):** Se realiza un conteo de los conectores causales en el texto que se encuentren clasificados de dicha forma en el archivo elaborado.

- **Incidencia de conectores lógicos (CNCLogic):** Se realiza un conteo de los conectores lógicos en el texto que se encuentren clasificados de dicha forma en el archivo elaborado.

- **Incidencia de conectores adversativos (CNCADC):** Se realiza un conteo de los conectores adversativos en el texto que se encuentren clasificados de dicha forma en el archivo elaborado.

- **Incidencia de conectores temporales (CNCTemp):** Se realiza un conteo de los conectores temporales en el texto que se encuentren clasificados de dicha forma en el archivo elaborado.

[9] Una versión online se encuentra en http://aplica.rae.es/grweb/cgi-bin/buscar.cgi
[10] La lista de conectores elaborada se encuentra en el Anexo 2

- **Incidencia de conectores aditivos (CNCAdd):** Se realiza un conteo de los conectores aditivos en el texto que se encuentren clasificados de dicha forma en el archivo elaborado.

6. Módulo de Complejidad Sintáctica

Se implementó la métrica de número de modificadores por sintagma nominal. Para ello se usó el analizador sintáctico de Freeling, el cual nos proporciona un árbol de sintaxis en donde cada uno de los nodos es una estructura sintáctica previamente especificada en un archivo de configuración.

- **Modificadores por sintagma nominal (SYNNP):** Se detectan los sintagmas nominales y se cuenta la media de modificadores en cada uno de estos. Un modificador es identificado como modificadores a los adjetivos, descriptivos y números.

7. Módulo de Densidad de Patrones Sintácticos

En este módulo se realiza el conteo de patrones sintácticos dentro del texto, tipos de palabras y tipo de sintagmas. Se implementaron las siguientes métricas:

Densidad de frases nominales (DRNP), Densidad de frases verbales (DRVP) y Densidad de negaciones (DRNEG). En estas 3 métricas se tomaron las definiciones de sintagma verbal y nominal especificadas por Freeling en su configuración para español.

8. Word Information

Las métricas contenidas en este módulo muestran la cantidad de palabras que existen en cada tipo de palabra (sustantivos, verbos, pronombres, etc.). Fueron implementadas 10 de 22 métricas de este grupo. Para la identificación de los tipos de palabras fue usado el PoS tagger[11] de Freeling.

- **Incidencia de sustantivos (WRDNOUN):** Consiste en realizar un conteo de los sustantivos que existen en el texto.

- **Incidencia de verbos (WRDVERB):** Consiste en realizar un conteo de los verbos que existen en el texto.

- **Incidencia de adjetivos (WRDADJ):** Consiste en realizar un conteo de los adjetivos que existen en el texto.

- **Incidencia de adverbios (WRDADV):** Consiste en realizar un conteo de los adverbios que existen en el texto.

- **Incidencia de pronombres (WRDPRO):** Consiste en realizar un conteo de los pronombres personales que existen en el texto.

- **Incidencia de pronombres personales primera persona singular (WRDPRP1s**): Consiste en realizar un conteo de los pronombres personales en primera persona singular que existen en el texto.

[11] PoS tagger (Part-of-Speech tagger) o también conocido como etiquetador gramatical, el cual te brinda información sobre la función que cumple una palabra en una oración.

- **Incidencia de pronombres personales primera persona plural (WRDPRP1p**): Consiste en realizar un conteo de los pronombres personales en primera persona plural que existen en el texto.

- **Incidencia de pronombres personales segunda persona (WRDPRP2):** Consiste en realizar un conteo de los pronombres personales en segunda persona que existen en el texto.

- **Incidencia de pronombres personales tercera persona singular (WRDPRP3s):** Consiste en realizar un conteo de los pronombres personales en tercera persona singular que existen en el texto.

- **Incidencia de pronombres personales tercera persona plural (WRDPRP3p):** Consiste en realizar un conteo de los pronombres personales en tercera persona plural que existen en el texto.

9. Legibilidad

Dentro de este paquete se implementó la fórmula de legibilidad de Flesh Fernandez, para lo cual se reutilizaron los cálculos hechos en el módulo de descriptivos quedando la fórmula:

Flesh = 206.84 - 60 * Media de sílabas por palabra - 102 * Media de palabras por oración

Consideraciones Finales

Se lograron implementar las métricas factibles dentro del proyecto, y estas nos sirvieron de base para hacer experimentación con el corpus recolectado. La mayor parte de la configuración en cuanto a la detección de estructuras y diccionarios esta por parte de Freeling, para lo cual se calibraron algunos parámetros para la detección de oraciones y estructuras.

CAPÍTULO 4

1 Corpus de textos en español

Dada la necesidad de tener un conjunto de textos que sean "naturalmente simples" y otro "naturalmente complejo" se optó por el género literario para la recopilación del corpus, estando este por cuentos y fábulas. Para el proyecto se define a un texto "naturalmente simple" o "naturalmente complejo" según el público al que está dirigido el texto. Por ejemplo, los textos simples han sido extraídos de fábulas y cuentos para niños, mientras que los textos complejos fueron extraídos de relatos dirigidos a adultos.

El corpus de textos fue recopilado de una sola fuente para los textos simples (o dirigidos a niños), pero se necesitaron de varias para los textos complejos. En la Tabla 7 se pueden observar las fuentes y el número de textos extraídos de cada uno.

Tabla 7. Distribución de corpus de textos

Fuente	Código	Número de textos
Cuentos cortos para niños[12]	C-NIÑO	50
Total Simples (niños)		**50**
Cuentos orientales[13]	C-ORIENTAL	7
Cuentos para reflexionar sobre la vida[14]	C-GALEON	11
Cuentos para adultos – El maestro cuenta cuentos[15]	C-ADULTOS	26
El vuelo del halcón[16]	C-VUELO	3
Cuentos para padres[17]	C-PADRES	3
Total Complejos (adultos)		**50**

Cada uno de los textos extraídos fue colocado inicialmente en un documento en MS Word para poder hacer uso del corrector ortográfico de dicha herramienta. Esto se realizó con el objetivo de eliminar posibles errores que pudiesen surgir luego al intentar extraer las métricas psicolingüísticas de legibilidad implementadas.

Como segundo paso se colocó solo el contenido de cada uno de los textos en un archivo de texto (un archivo por cada texto) para que estos puedan ser procesados con mayor facilidad por la herramienta de cálculo de métricas y se asemeje mejor al formato de entrada que recibirá la aplicación final.

Finalmente, se elaboró una versión más por cada uno de los textos que consta de un archivo XML con la estructura mostrada en la Figura 7. Como se observa, este archivo no cuenta solo con el contenido del texto, sino también con datos adicionales como el

[12] Revisada el 21 de Setiembre de 2013. Web: http://cuentoscortosparaniños.org/
[13] Revisada el 21 de Setiembre de 2013. Web: http://www.terapiapsico-corporal.com/2013/04/cuentos-breves-orientales-sabiduria-terapia-sufis-zen.html
[14] Revisada el 21 de Setiembre de 2013. Web: http://www.galeon.com/jlgarcia/11cuentos/11cuentos.htm
[15] Revisada el 21 de Setiembre de 2013. Web: https://elmaestrocuentacuentos.wikispaces.com/Cuentos+para+adultos
[16] Revisada el 21 de Setiembre de 2013. Web: http://isabelquiroz.wordpress.com/el-vuelo-del-halcon/
[17] Revisada el 21 de Setiembre de 2013. Web: http://www.encuentos.com/cuentos-para-padres/fotos/

título, complejidad y referencia. Estos archivos se generaron con el objetivo que su estructura pueda servir para proyectos futuros donde se necesiten que el corpus de textos tenga mayores anotaciones como número de oraciones, estructuras sintácticas, entre otras.

```
<titulo>Los gemelos y las hormigas</titulo>
<complejidad>simple</complejidad>
<referencia>http://cuentoscortosparaniños.org/</referencia>
<contenido>
Después de un gran almuerzo
en una ciudad lejana, una familia
celebraba la graduación de uno
de sus hijos.
...
Los gemelos dijeron juntos:
A partir de ahora nadie las
molestará y les pondremos
granos de arroz y algunas migas
de pan para que tengan mucha
comida para todos en invierno.
</contenido>
```

Figura 7 Ejemplo de archivo XML del corpus

2 Modelo de clasificación de complejidad de textos en español

Una vez terminado el proceso de recopilación del corpus, se procedió a hacer un cálculo de las métricas para cada uno de los textos. Se elaboró un consolidado de los resultados que se muestra en la Tabla 8 para poder verificar que realmente los textos cumplan con la división planteada inicialmente: “naturalmente simple” y “naturalmente complejo”. Dicha tabla está agrupada por los grupos que se encuentran en Coh-Metrix 3.0 (Inglés), de la misma forma los códigos mostrados son equivalentes a los de la herramienta antes mencionada.

Tabla 8 Resumen de métricas para el corpus

Grupo o módulo	Métrica	Simple	Complejo
Descriptives	DESPC	224	821
	DESSC	907	2432
	DESWC	16552	33326
	DESPL	7360.64	12069.99
	DESSL	1022.54	739.72
	DESWLsy	98.51	101.21
	DESWLlt	220.21	223.91
Referential Cohesion	CRFNO1	16.67	6.83
	CRFAO1	29.33	16.15
	CRFSO1	19.55	8.01
	CRFNOa	16.56	6.84
	CRFAOa	28.31	15.39
	CRFSOa	19.01	7.79
	CRFCWO1	5.52	4.71
	CRFCWOa	6.31	4.81
Lexical Diversity	LDTTRc	28.03	26.07
	LDTTRa	24.9	22.48

Connectives	CNCAll	1	1.98
	CNCCaus	0.09	0.11
	CNCLogic	0.59	1.17
	CNCADC	0.18	0.27
	CNCTemp	0.14	0.34
	CNCAdd	0.6	1.25
Syntactic Complexity	SYNNP	32.35	36.74
Syntactic Pattern Density	DRVP	852	2187
	DRNEG	222	511
Word Information	WRDNOUN	3.42	6.83
	WRDVERB	3.59	6.99
	WRDADJ	0.77	1.6
	WRDPRO	0.54	1.16
	WRDADV	1.04	2.04
	WRDPRP1s	0.08	0.34
	WRDPRP1p	0.03	0.05
	WRDPRP2	0.04	0.18
	WRDPRP3s	0.27	0.53
	WRDPRP3p	0.11	0.06
Readability	RDFFGL	83.77	79.09

Con esta tabla se puede notar claramente que existe una diferencia entre los textos simples y los complejos. Por ejemplo, en cuanto a número de párrafos los textos simples tienen un menor número comparado a los complejos. De la misma forma, si se analiza el grupo *Referential Cohesion* se puede notar que los textos complejos tienen menores valores que los simples, lo cual indica que efectivamente estos son más complejos ya que este grupo de métricas mide cuan unidas están las ideas entre ellas y un valor menor muestra que no existe mucha relación. Por último, al analizar la métrica en *Readability* usado para medir la legibilidad de un texto se puede notar que los textos complejos obtienen un menor valor comparado a los textos simples.

Por tanto, en base a los datos mostrados en la Tabla 8 se puede notar que dicho corpus está bien dividido según su complejidad.

Para hallar el modelo de clasificación de complejidad de textos se procedió a experimentar utilizando diversos algoritmos de aprendizaje de máquina de tipo supervisado. Estos algoritmos fueron proveídos por la herramienta WEKA.

Dado que no solo debía obtenerse el modelo de clasificación, sino una herramienta que pueda clasificar textos en base a un entrenamiento previo, se implementó una aplicación utilizando la versión de desarrollo de WEKA, que es usada como una librería del proyecto.

Inicialmente, se calculó las métricas para cada texto y se creó un archivo de texto con los valores de cada métrica (un archivo por cada texto) tal como se muestra en la Figura 8.

```
DESPC 10.0
DESSC 18.0
DESWC 347.0
DESPL 155.13008877856174
DESSL 19.27777777777778
...
WRDPRP1s 0.001
WRDPRP1p 0.001
WRDPRP2 0.0
WRDPRP3s 0.001
WRDPRP3p 0.009
RDFFGL 88.22996498101085
```

Figura 8 Ejemplo de métricas calculadas

Con las métricas para cada texto calculadas, se procedió a elaborar un único archivo de formato ARFF, formato que soporta WEKA para el análisis de la data. La estructura de dicho archivo se muestra en la Figura 9.

```
@RELATION corpus

@ATTRIBUTE CRFSO1 REAL
@ATTRIBUTE CRFANP1 REAL
...
@ATTRIBUTE DESPC REAL
@ATTRIBUTE class {simple,complejo}

@DATA
0.117647,0.000000, ... ,0.014000,10.000000,simple
...
0.500000,0.000000, ... ,0.054000,23.000000,complejo
```

Figura 9 Ejemplo de archivo con formato ARFF

Como se observa en la Figura 9, se tiene una lista de los atributos y el tipo de dato de cada uno. Además, se añade como un atributo más la clase y sus diferentes valores. Luego de listar los atributos se procede a colocar una fila por cada texto con los valores de cada uno de los atributos.

Teniendo como base la estructura del archivo, se colocarán como atributos la lista de las métricas implementadas y como clases *simple* y *complejo*. Se tendrán también cincuenta filas con clasificación simple y cincuenta con clase compleja dado que esta es la división realizada en el corpus.

Una vez generado dicho archivo se procede a realizar el experimento, el cual será ejecutado mediante el método 10-fold cross-validation para cada uno de los algoritmos de aprendizaje de máquina de tipo supervisado que provea la librería. Este método implica dividir el corpus en diez partes iguales y tomar siempre una décima parte para comprobar la correctitud del algoritmo, mientras que lo demás servirá para entrenar al modelo, esto se realizará diez veces y se hará una última ejecución del algoritmo con el mejor modelo obtenido, este método es brindado por WEKA. El experimento consistió en entrenar a todos los algoritmos involucrados y obtener un ranking por

precisión de estos. En la Tabla 9 se puede observar a los tres mejores algoritmos y dos más que serán usados como *baseline* para analizar los resultados.

Tabla 9 Resultado Experimento

Algoritmo	Precision	Recall	F-Measure
SMO	0.9	0.9	0.9
SimpleLogistic	0.88	0.88	0.88
LMT	0.88	0.88	0.88
OneR	0.82	0.8	0.8
ZeroR	0.25	0.5	0.33

El mejor modelo de clasificación utiliza el algoritmo SMO[18] con una precisión de 90% para el corpus de textos y las métricas implementadas.

A pesar de obtenerse una muy buena precisión se puede notar también que el algoritmo OneR, que toma decisiones en base a un solo atributo[19] de todos los enviados al modelo, obtiene una precisión de 80%. Esto puede deberse principalmente a dos situaciones: la tarea de clasificación de textos en el corpus recolectado es muy sencilla de realizar o las métricas brindan información bastante relevante sobre los textos, lo que facilita la clasificación de estos.

Por otro lado, el algoritmo ZeroR, que toma decisiones tomando solo como referencia a la clase que posee un mayor número de instancias (textos), obtuvo un 25% de precisión muy por debajo de SMO, esto dado que la distribución de textos por clase está balanceada. Esto muestra que para el corpus dado, el conocer solo su clase no es suficiente para poder obtener un buen resultado de clasificación en el modelo.

El detalle del desempeño de SMO está en la Tabla 10, la cual muestra que de los 50 textos simples, 47 fueron clasificados correctamente con el modelo y 3 de ellos fueron clasificados como complejos; de manera similar, 43 textos complejos clasificados correctamente por el modelo y los restantes como simples.

Tabla 10 Matriz de confusión de SMO

Simple	Complejo	Classified as
47	3	Simple
7	43	Complejo

Dado que se logró cumplir con la meta de al menos 60% de precisión, se realizaron dos experimentos más para poder analizar el comportamiento de los algoritmos frente a otro corpus de textos. Este nuevo corpus fue obtenido con el mismo método especificado en el resultado 3[20], pero esta vez los textos eran dirigidos a estudiantes de español como lengua extranjera y sus clases eran equivalentes a los niveles de enseñanza que se tienen para su enseñanza: básico, intermedio y avanzado.

Con este nuevo corpus, de 31 textos, se realizó el experimento con los algoritmos de aprendizaje de maquina obteniendo los resultados mostrados en la Tabla 11.

[18] SMO: Sequential Minimal Optimization, es una optimización de Support Vector Machines
[19] En este experimento el atributo que eligió fue "Incidencia de todos los conectores" (CNCAll)
[20] El resultado 3 consta del corpus de textos (50 fábulas para niños y 50 cuentos para adultos)

Tabla 11 Resultado Experimento con corpus de 3 clases

Algoritmo	Precision	Recall	F-Measure
FilteredClassifier	0.72	0.77	0.73
AdaBoostM1	0.7	0.74	0.69
DecisionTable	0.7	0.74	0.69
OneR	0.69	0.71	0.68
ZeroR	0.27	0.52	0.35

Los resultados del experimento muestran que se obtuvo una menor precisión para clasificar textos al agregar una clase más y el mejor modelo utiliza el algoritmo FilteredClassifier con precisión de 72%. Sin embargo, el algoritmo OneR[21] sigue obteniendo una precisión que no es considerablemente menor al mejor modelo. En la Tabla 12 se muestra la matriz de confusión de FilteredClassifier en la que se observa que de los 3 textos avanzados que se tienen no se pudo clasificar correctamente ninguno, mientras que para la clase intermedio se clasificó la totalidad de estos como tal.

Tabla 12 Matriz de confusión FilteredClassifier

basico	intermedio	avanzado	Classified as
8	4	0	basico
0	16	0	intermedio
1	2	0	avanzado

La distribución del corpus con tres clases no es uniforme, es decir, la cantidad de textos por clase es muy diferente entre cada una de estas, lo cual causa que el modelo no pueda generalizar correctamente y clasificar exitosamente a los textos avanzados. Esto lleva a realizar un segundo experimento con este corpus, pero en este caso se quitó los tres textos avanzados obteniendo los resultados de la Tabla 13.

Tabla 13 Resultado Experimento sin textos avanzados

Algoritmo	Precision	Recall	F-Measure
Logistic	0.84	0.82	0.82
MultiClassClassifier	0.84	0.82	0.82
MultiClassClassifierUpdateable	0.8	0.79	0.79
OneR	0.71	0.71	0.71
ZeroR	0.33	0.57	0.42

Se ve que la precisión del mejor modelo de clasificación mejoró considerablemente con respecto al anterior experimento. Esta vez el mejor modelo de clasificación obtuvo un 84% de precisión mientras que el algoritmo OneR solo 71%. Al observar la matriz de confusión de Logistic (Tabla 14), usado en el mejor modelo para el experimento, se puede notar que a pesar de que el número de textos intermedios clasificados correctamente disminuyó, en total fueron 23 textos clasificados correctamente, entre básicos e intermedios.

[21] El algoritmo OneR elige la métrica "Incidencia de sustantivos" para realizar la clasificación

Tabla 14 Matriz de confusión Logistic

basico	intermedio	avanzado	Classified as
11	1	0	basico
4	12	0	intermedio
0	0	0	avanzado

3 Consideraciones finales

Estos experimentos demuestran así que el poseer una distribución de textos por clase no uniforme afectará a la precisión total del modelo de clasificación; es decir, que si se desea agregar una clase más al corpus inicialmente planteado (simple y complejo), el conjunto de textos con dicha clase nueva deberá ser similar en cantidad a los dos anteriores.

CAPÍTULO 5

1 Servicio web de clasificación automática de complejidad de textos

Con el modelo ya generado y el motor de cálculo de métricas terminado, se paso a desarrollar la herramienta web de clasificación automática. Para ello, se agrupo el clasificador y el motor de cálculo de métricas en un servidor HTTP (Spark), el cual por medio de llamadas POST recibe los textos a clasificar. Una vez terminado el procesamiento, se retornan tanto la clase como las métricas utilizadas para la clasificación.
Los datos calculados por el servidor HTTP de métricas y clasificación, son usados por la interfaz web, desarrollada utilizando el framework Ruby on Rails para las llamadas http y generación de reportes y Twitter Bootstrap, para los estilos y diseño final de los informes.

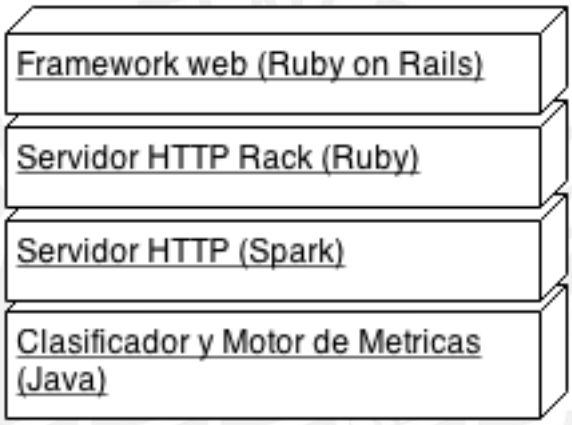


Figura 10 Arquitectura herramienta web

Para el uso de la herramienta se tiene un formulario que solo contiene una entrada de texto, en la cual se inserta o copia el texto que se desea evaluar. Después de ello, se envía esta información para su procesamiento y se retorna un informe sobre la predicción de la clase y las métricas utilizadas para esta clasificación.

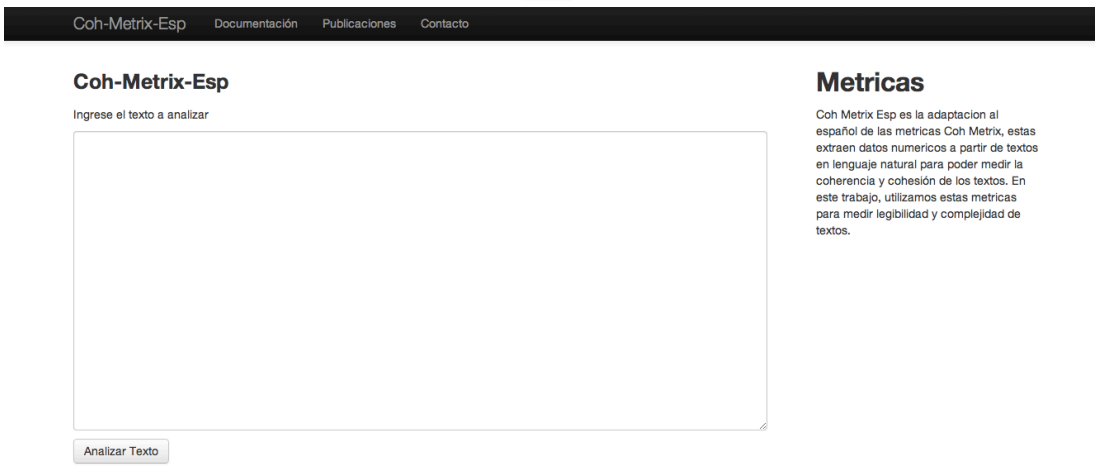
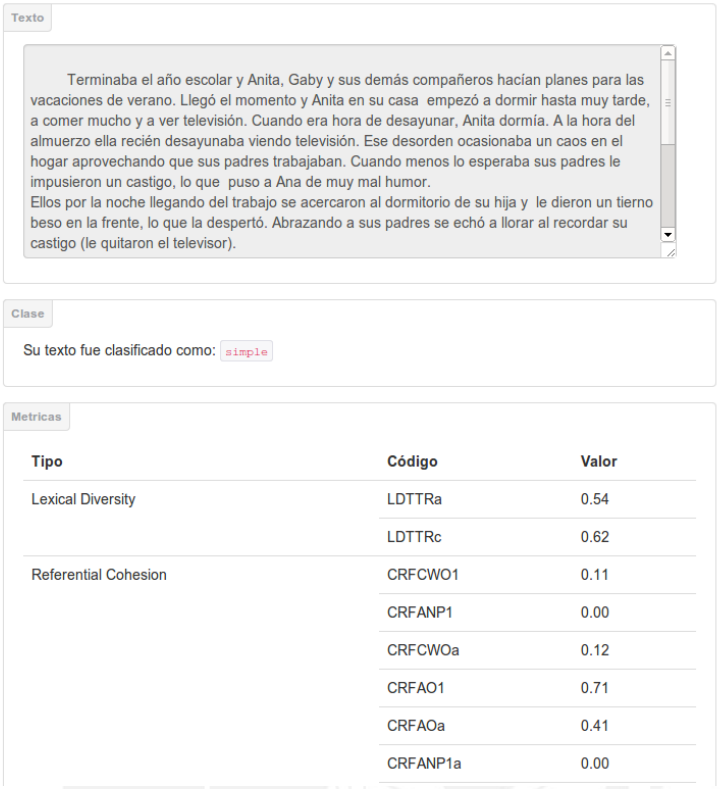


Figura 11 Interfaz web de Coh-Metrix-Esp

Para el informe de las métricas calculadas, se dividió en las mismas categorías y códigos que Coh-Metrix 3.0.



Tipo	Código	Valor
Lexical Diversity	LDTTRa	0.54
	LDTTRc	0.62
Referential Cohesion	CRFCWO1	0.11
	CRFANP1	0.00
	CRFCWOa	0.12
	CRFAO1	0.71
	CRFAOa	0.41
	CRFANP1a	0.00

Figura 12 Resultado de ejecución Coh-Metrix-Esp

Por la forma en la que se desarrollo la herramienta. se cuenta con la posibilidad de escalar sin problemas el número de clasificadores y extractores de métricas. Es decir, las instancias de procesamiento están divididas por lo que se puede utilizar técnicas de balance de carga y soportar una gran cantidad de conexiones al mismo tiempo. El código está disponible en: https://github.com/andreqi/coh-metrix-esp-front-en

CAPÍTULO 6

1 Conclusiones

En el desarrollo de la herramienta de análisis y clasificación de complejidad de textos en español exploramos las herramientas existentes para el soporte del desarrollo de herramientas de procesamiento de lenguaje natural, encontrando librerías muy completas pero herramientas desarrolladas limitadas.

La facilidad que nos dieron las herramientas utilizadas en el desarrollo de Coh-Metrix-Esp aligeró la carga en la recopilación de recursos, lo cual dejó tiempo adicional para el desarrollo del análisis de las métricas, creación del corpus, clasificadores y herramienta web.

Adicionalmente, dentro del análisis realizado con las métricas sobre el corpus recopilado, se pudo apreciar la utilidad de las métricas adaptadas, tal como se observó en el Capítulo 4, las cuales ya delimitaban notablemente las dos clases del corpus de textos. Además, la precisión de los clasificadores basados en estas métricas supera nuestra hipótesis inicial del 60% obteniendo un valor de 90% para el corpus de textos recopilado.

Al obtener buenos resultados con el corpus recopilado y las clases definidas, se pasó a hacer un análisis adicional, considerando tres clases con un corpus diferente, para lo cual se recopilaron textos para alumnos que estudian español como segundo idioma. Este corpus de 31 textos, no fue suficiente para la generalización de esta clase, por lo que los resultados de los clasificadores basados en las métricas implementadas y los corpus de la nueva clase adicional bajaron un 40%. Esta baja fue a causa de la reducida cantidad de textos de la clase 'avanzado'. Es por ello que si se desea agregar una clase adicional para el clasificador, se necesita mantener la proporción en cada una de las clases, ya que la no uniformidad afecta considerablemente a la precisión del modelo de clasificación.

2 Trabajos futuros

Debido al alcance del proyecto se pueden realizar las siguientes variaciones:

- *Cambiar el género del corpus de textos:*

  El cambiar el género del corpus podría brindar una mejor generalización de los textos. Por ejemplo, si se eligiera género periodístico, se tendría una mayor cantidad de temas lo que ayudaría al modelo de clasificación a obtener mejores respuestas para una mayor cantidad de textos.

- *Ampliar la cantidad de textos por clase:*

  Con esto se podría probar de una mejor manera la precisión del modelo y a pesar que la precisión podría bajar en comparación a la de este proyecto, esto estaría sustentado en una mayor cantidad de textos a analizar, y se podrían realizar mayores experimentos dado que podrían incluirse textos simples cercanos a los complejos y viceversa.

- *Agregar clases o niveles de clasificación al corpus de textos:*

  Como se ha observado en los experimentos, el agregar niveles de clasificación (o clases) al corpus puede disminuir la precisión del modelo de clasificación si no está uniformemente distribuido. Sin embargo, si se puede conseguir esto, ofrecería mayores beneficios, dado que podrían llegarse a tener tantas clases como se pueda. Por ejemplo, una clase por cada grado de primaria escolar, secundaria escolar, etc.

- *Implementar más métricas psicolingüísticas de legibilidad:*

  Dentro del proyecto, se llegaron a implementar 48 de 106 métricas de Coh-Metrix 3.0, por lo que queda aún gran parte que implementar. Muchas de estas métricas fueron dejadas de lado por la complejidad, los recursos necesarios o el alcance del proyecto.

  El desarrollar las métricas restantes o proponer métricas adicionales contribuye a un análisis más exacto de la complejidad de textos, lo cual genera una mayor cantidad de aplicaciones y herramientas por desarrollar, no solo ligadas a la clasificación de textos. Con más métricas se pueden realizar análisis de aprendizaje no supervisado sobre textos sin clasificación previa como corpus.

- *Mejorar la escalabilidad de las métricas de cohesión referencial:*

  En la implementación de estas métricas, se necesita la comparación de todas las oraciones en pares. Para textos con más de 10000 oraciones, esto es un problema, ya que dentro de la implementación realizada en el presente proyecto, se hace un comprobación por fuerza bruta.

  Se puede disminuir la complejidad de estos cálculos como trabajo futuro, para que estas métricas puedan ser aplicadas a textos voluminosos.

- *Realizar un estudio de selección de métricas para enviar al clasificador:*

  Dentro del presente proyecto se utilizan todas las métricas como entrada de clasificación, lo cual puede ser un problema, ya que algunas métricas pueden introducir ruido y no determinen la pertenencia a una clase. Por lo que se deja como trabajo futuro la evaluación de cada una de las métricas implementadas, en cuanto a su importancia en el clasificador.

  El estudio de cada una de estas métricas puede dar resultados en cuanto a la relación de métricas para determinados corpus. Esto puede ser de utilidad en el caso de la identificación de géneros literarios y estilos.

Referencias bibliográficas

Centro Cervantes
2012 http://cvc.cervantes.es/lengua/anuario/anuario_12/i_cervantes/p01.htm , Anuario

CHISWELL, Ian; HODGES, Wilfrid
2007 Mathematical Logic. Oxford, United Kindom: Oxford University Press

Coh-Metrix 3.0
2011 cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html , Documentación

CROSSLEY, Scoot; MCCARTHY, Philipl; Louwerse, Max; MCNAMARA, Denielle
2007 A Linguistic Analysis of Simplified and Authentic Texts, The Modern Language Journal, USA, 2007, Vol 91, No.7 pp. 15-30

Examen PISA
2007, http://pisa-sq.acer.edu.au/showQuestion.php?testId=2292&questionId=12 Sample Questions. Revisada en Agosto de 2013

FERNÁNDEZ Huerta, J
1959 Medidas sencillas de lecturabilidad. Consigna,

FISHER, Douglas; FREY, Nancy; LAPP, Diane
2012 Text Complexity: Raising Rigor in Reading. Delaware, USA: International Reading Asociation.

Freeling
2012 http://nlp.lsi.upc.edu/freeling, Documentación. Revisada en Agosto de 2013

GRAESSER,Arthur C.;MCNAMARA, Danielle S. ; KULIKOWICH, Jonna M.
2011 "Coh-Metrix:Providing Multilevel Analyses of Text Characteristics". Educational Researcher. USA, 2011, Vol.40, No. 5, pp. 223-234

GRAESSER, Arthur C. ; MCNAMARA, Danielle S. ; LOUWERSE, Max M ; CAI,Zhiqiang
2004 "Coh-Metrix: Analysis of text on cohesion and language" Behavior Research Methods, Instruments & Computers, 2011

GRAESSER,Arthur C. ;PETSCHONEK, Sarah
2005 "Automated System that Analyze Text and Discourse: QUAID, Coh-Metrix and AutorTutor" University of Memphis, Department of Psychology

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome
2010 The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Stanford, California, USA: Springer

Inflesz
2007 http://legibilidad.com/home/acercade.html#inflesz, Página principal. Revisada en Agosto de 2013

BISHOP, Christopher
2006 Pattern Recognition and Machine Learning. New York , USA: Springer

JURAFSKY, Daniel; MARTIN, James H.
2007, "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition"

Legibilidad
2007 http://legibilidad.com/home/acercade.html#legibilidad. Revisada en Agosto de 2013.

MANNING, Christopher
2000 Fundations of Statistical Natural Language Processing, Massachusetts, USA: MIT

MCNAMARA, Danielle; CROSSLEY, Scott
2010 "Cohesion, Coherence, and Expert Evaluations of Writing Proficiency". Proceedings of the Anual Meeting of the Cognitive Science Society. USA, No. 12

Ministerio de educación, Gobierno del Perú
2013 Evaluación Censal de Estudiantes 2013

Organization for Economic Co-operation and Development
2012 PISA 2012 Results: What Students Know and Can Do , "Student Performance in Reading, mathematics and science" , Vol 1

PETERSEN, Sarah
2007 Natural Language Processing Tools for Reading Level Assessment and Text Simplification / Doctorado en Procesamiento de lenguaje natural / Doctorado : Universidad de Washington / Escuela de Graduados

PETERSEN, Sarah; OSTENDORF, Mari
2008 A machine learning approach to reading level assessment. Computer Speech and Language, USA, 2009, No 23, pp. 89-106

PROJECT MANAGEMENT INSTITUTE (PMI)
2013 Project Management Body Of Knowledge (PMBOK). 5ta Edición. Pensilvania.

SCARTON, Carolina
2009, Avaliação da Inteligibilidade de Textos em Português: uma aplicação na área de simplificação de textos para o público infantil / Iniciación científica / Bachillerato : São Paulo:Universidade São Paulo/ Instituto de Ciências Matemáticas e de computação / Iniciação Científica

SCARTON, Carolina
2010 Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. Universidade de Sao Paulo. Instituto de Ciências Matemáticas e de computação

SCHWARM, Sarah; OSTENDORF, Mari
2005 Readling level assessment Using Support Vector machines and Statistical Language Models. Proceedings of the Annual Meeting of the ACL, USA, No. 43

Society for human resource management
2010 Are they ready to work?: Employers Prespectives on the Basic Knowledge and Applied Skills of New Entrants to the 21st Century US Workforce

Warck, José Antonio
2005, Segmentación de palabras en sílabas, Tesis de licenciatura

WEITZENFELD, Alfredo
2004 Ingeniería de software orientada a objetos con UML, Java e Internet. Primera edición. México: Thomson.

WEKA
2013 http://weka.wikispaces.com, Documentación. Revisada en Septiembre de 2013