Estudio de la complejidad del Español para la simplificación textual

Randall Araya-Camposa[1], Paula Estrella[2], José Arguedas-Castillo[3], Walter Alvarez-Grijalba[4]

Araya-Camposa, R; Estrella, P; Arguedas-Castillo, J; Alvarez-Grijalba, W. Estudio de la complejidad del Español para la simplificación textual. *Tecnología en Marcha*. Vol. 33, especial Movilidad estudiantil. Pág 45-63.

https://doi.org/10.18845/tm.v33i7.5478

1 Escuela de Ingeniería en Computación. Centro Académico de Alajuela. Instituto Tecnológico de Costa Rica, Alajuela, Costa Rica. raraya@ic-itcr.ac.cr
2 Facultad de Lenguas y FaMAF. Universidad Nacional de Córdoba Córdoba, Argentina. pestrella@famaf.unc.edu.ar
3 Escuela de Ingeniería en Computación. Centro Académico de Alajuela. Instituto Tecnológico de Costa Rica. Alajuela, Costa Rica. joarguedas@ic-itcr.ac.cr
4 Escuela de Ingeniería en Computación. Centro Académico de Alajuela. Instituto Tecnológico de Costa Rica. Alajuela, Costa Rica. walvarez@ic-itcr.ac.cr

Palabras claves

Lecturabilidad; métricas de complejidad; simplificación textual; evaluación.

Resumen

La mayoría de los trabajos en el área de la simplificación textual se realizan sobre el idioma inglés por contar con más recursos lingüísticos y sobre el género periodístico. Sin embargo, por nuestro contexto, en este trabajo nos concentramos en estudiar y automatizar las métricas existentes para medir complejidad léxica para el español, como paso previo a la identificación de frases complejas y su subsiguiente simplificación. Otro aspecto novedoso de este trabajo es la utilización de corpus relacionados a los derechos humanos, concretamente de la Organización para las Naciones Unidas y del Alto Comisionado de las Naciones Unidas para los Refugiados. Los aportes más significativos son: la creación de una herramienta de código abierto, que genera un informe sobre la complejidad de un texto dado con el fin de dar soporte a quien esté interesado en simplificar ese texto, y la propuesta de una nueva métrica para medir la complejidad de manera multifacética. Los resultados obtenidos en los distintos experimentos realizados son prometedores y en muchos casos confirman las hipótesis planteadas.

Introducción

El décimo Objetivo de Desarrollo Sustentable de la Organización de las Naciones Unidas (ONU o UN por sus siglas en inglés) es *Reducir inequidades*[5]. Aunque parezca tan ajeno a nosotros o tan ambicioso, este objetivo incluye aspectos de lo cotidiano, como por ejemplo el acceso a la información, el cual es un derecho humano básico en una democracia moderna [1] y, a pesar de la masividad de información disponible en diversos formatos y medios, aún existen personas que no logran acceder a ella por falta de comprensión de la misma. Algunos de los motivos son la generación de textos no apropiados para ciertas edades (por ejemplo, un niño de 10 años seguramente no comprenda un informe médico), la presencia de alguna patología (por ejemplo, enfermedades que afecten la memoria, dislexia, afasia, etc) o un nivel de alfabetización no adecuada. En el peor de los casos, esto resulta en personas a las que se les violenten sus derechos ya que no pueden comprender los textos (leyes, reglamentos, constituciones, etc) que los protegen.

Una posible solución al acceso a la información es la adaptación de los textos a la audiencia que van dirigidos. Por ejemplo, un cuento para niños en edad pre-escolar debe tener estructuras sintácticas y un vocabulario sencillos, mientras que un artículo científico debe tener un vocabulario técnico y específico del área. Esta adaptación puede realizarse en forma manual o (semi-)automática y tiene como fin último mejorar la lecturabilidad de los textos. Aunque este término *lecturabilidad* no existe formalmente en el diccionario de la Real Academia Española, en este trabajo hemos decidido utilizarlo como un calco de la palabra en inglés *readability*, cuyo significado es el índice de facilidad que tiene un texto para ser comprendido por los lectores. De esta definición se desprende que la lecturabilidad es inversamente proporcional a la complejidad textual, es decir, necesitamos reducir la complejidad de los textos para ganar lecturabilidad[6]

La adaptación en forma automática (o semi-automática) es denominada *simplificación textual* (ST) y es una rama del Procesamiento del Lenguaje Natural (PLN) que tiene como objetivo

[5] Más información disponible en https://www.ar.undp.org/content/argentina/es/home/sustainable-development-goals.html
[6] Más información disponible en http://www.cuadernointercultural.com/lecturabilidadlegibilidad-comprension-lectora/

convertir un texto complejo en un texto más simple preservando su significado e idea principal, así aumentando su lecturabilidad. La ST puede suceder a distintos niveles: a nivel léxico (identificando y reemplazando ciertas palabras por sinónimos más comprensibles), sintáctico (descomponiendo estructuras compuestas de una frase en una o más estructuras simples), semántico (identificando y explicando conceptos abstractos y/o ambiguos) o discursivo (explicitando toda información que pueda aparecer tácitamente y/o referenciando a otros elementos según las reglas del idioma). En esta forma de adaptar textos es necesario combinar diversas áreas del PLN para las cuales no siempre se cuenta con los recursos necesarios. De hecho, uno de los recursos más escasos son los corpus anotados para tareas como resolución de anáforas, identificación de palabras complejas o para aprendizaje automático (supervisado). Esta es la principal limitación para el desarrollo de sistemas de ST completamente automáticos y lo suficientemente generales como para aplicarse a textos de cualquier género en un idioma.

Una etapa previa a la ST es la identificación de las partes de un texto que afectan negativamente su lecturabilid. Esta tarea ha cobrado importancia en los últimos años a tal punto que la prestigiosa SemEval (International Workshop on Semantic Evaluation)[7] ha incluido una tarea específica de identificación de palabras complejas apuntando a la simplificación léxica por primera vez en 2016 [2] y al momento de escribir este artículo se encuentra en curso la segunda edición de esta tarea[8].

Como describiremos en la siguiente sección, la mayoría de los trabajos en este área se realizan sobre el idioma inglés por contar con más recursos lingüísticos y sobre el género periodístico. Sin embargo, por nuestro contexto, en este trabajo nos concentramos en estudiar y automatizar las métricas existentes para medir complejidad léxica para el español, como paso previo a la identificación de frases complejas y su subsiguiente simplificación. Otro aspecto novedoso de este trabajo es la utilización de corpus relacionados a los derechos humanos, concretamente de la Organización para las Naciones Unidas (ONU) y del Alto Comisionado de las Naciones Unidas para los Refugiados (ACNUR).

Este artículo se organiza de la siguiente manera: la Sección II describe los trabajos más relevantes en el área de interés; en la Sección III detallamos las herramientas y recursos utilizados en este estudio; en la Sección IV se describe la metodología empleada en los experimentos detallados en la Sección V; en la Sección VI se contrastan los resultados obtenidos automáticamente en la sección anterior con los obtenidos con anotadores humanos y la Sección VII ilustra la simplificación textual automática para el caso puntual de los números romanos; concluimos este trabajo con reflexiones y trabajo futuro en la Sección VIII.

Trabajos relacionados

Los primeros trabajos en el área de ST automática, como [3], [4] y [5], estaban enfocados a preprocesar el texto para otra aplicación del PLN. Sin embargo, estos trabajos son la base para el primer sistema de ST desarrollado para facilitar la comprensión lectora a personas con afasia en el marco del proyecto *Syntactic simplification in PSET* [6]. Este proyecto abordó la simplificación sintáctica en artículos periodísticos en inglés, basándose en un conjunto de reglas creadas manualmente para la re-escritura de cláusulas coordinadas y el cambio de voz pasiva a voz activa, y en WordNet [7] para la sustitución de palabras categorizadas como díficiles de comprender según la base de datos *Oxford Psycholinguistic Database* [8].

7 https://en.wikipedia.org/wiki/SemEval
8 Detalles de la campaña actual disponibles en https://sites.google.com/view/cwisharedtask2018/home

El trabajo sobre el idioma inglés se vio muy beneficiado con la creación de *SimpleWikipedia*[9], que es una versión reducida de Wikipedia y escrita siguiendo reglas de producción de contenidos para que resulte fácil de comprender. Esto ha permitido crear recursos como corpus paralelos "inglés estándar" – "inglés sencillo" [9], que luego son empleados en la creación de sistemas de ST. Incluso es posible emplear técnicas que demandan un gran volumen de datos, por ejemplo, los siguiente trabajos han abordado la ST como una tarea de traducción tomando inglés complejo como fuente y generando inglés sencillo.

En [10] los autores se inspiran en la traducción automática estadística sintáctica [11] para realizar operaciones de borrado, separación de una oración en varias oraciones, reordenamiento y sustitución, las cuales en conjunto forman el modelo de traducción aprendido a partir de un corpus paralelo de más de 100 mil pares de frases extraídas de Wikipedia Simple y Wikipedia. Para evaluar su sistema utilizan la medida Flesch Reading Ease [12] y la perplejidad del modelo de lenguaje resultante del entrenamiento. Luego en [13], se agrega como recurso el historial de revisión de los artículos en Wikipedia Simple a los fines de aprender las partes de un artículo de Wikipedia (estándar) que deben ser simplificados. El formalismo empleado en este caso son las gramáticas quasi-síncronas que son la base de la generación el texto simplificado. En [14] se adopta un enfoque puramente estadístico, utilizando un corpus similar a los anteriores para entrenar un sistema de traducción automática por frases con la herramienta Moses [15]. Como esta herramienta no tiene ninguna característica enfocada a la ST, los autores generan las ***n*** mejores traducciones (lo cual sí es provisto por Moses) y heurísticamente seleccionan la que diverja más del texto fuente; la hipótesis es que mientras más diversa la traducción (inglés sencillo) al texto original (inglés estándar) más operaciones de simplificación se han producido. Los resultados de su evaluación no muestran mejoras significativas respecto de [10] por ejemplo.

Para el caso de otros idiomas que no cuentan con una fuente de textos paralelos como Wikipedia Simple, estos deben generarse manualmente o recolectarse de manera semiautomática. Los idiomas de interés regional son el español y el portugués (de Brasil), por lo que se describe a continuación el trabajo realizado en el marco de dos proyectos sobresalientes: PorSimples [16] y Simplext [17].

El proyecto PorSimples se ejecutó en Brasil entre los años 2007 – 2010 y generó una serie de recursos y herramientas para el Portugués de Brasil con el objetivo de beneficiar tanto a la población con bajo nivel de alfabetización como a los autores que deseen producir textos para estas audiencias. Los recursos y herramientas, detallados en [18], son:

- Un manual para la ST del Portugués (BR), donde se dan recomendaciones sobre cómo tratar ciertos fenómenos sintácticos.
- 9 corpora de 2 géneros distintos (noticias generales y artículos de ciencia popular); los artículos fueron simplificados manualmente por lingüistas especializados en ST.
- Un diccionario de palabras simples que contiene palabras frecuentes entre los jóvenes y en textos de noticias para niños, entre otras; este diccionario se utiliza para la tarea de identificación de palabras complejas.
- La herramienta de apoyo a la creación de textos simplificados *SIMPLIFICA*.
- Un sistema de asistencia a la lectura, llamado *FACILITA*, el cual aplica algunas estrategias de resumen y simplificación automática para ayudar a lectores en la web.

[9] https://simple.wikipedia.org/

- Una herramienta de adaptación de contenidos, denominada *Educational FACILITA* que resalta contenido específico (como entidades nombradas, relaciones entre verbos, etc) que ayudan al lector a comprender mejor el texto.

El proyecto *Simplext*, a diferencia de PorSimples, tiene como objetivo ayudar a personas con algún impedimento cognitivo, como por ejemplo afasia. Este proyecto generó los siguiente productos:

- Un corpus paralelo del español, simplificando manualmente alrededor de 200 noticias cortas.
- Un módulo de simplificación léxica denominado *LexSis* que utiliza recursos libres para aplicar técnicas de sustitución léxica a base de un modelo vectorial basado en palabras, la frecuencia de las palabras y la longitud de la palabras [19].
- Un módulo de simplificación sintáctica que realiza algunas operaciones de transformación al nivel de la oración, según lo observado en el corpus paralelo, y algunas transformaciones basadas en reglas sobre frases idiomáticas y expresiones que no pueden simplifacarse por mera sustitución de palabras, como por ejemplo las expresiones numéricas [20].

Una parte del proyecto se dedicó a la simplificación de expresiones numéricas, como cifras, porcentajes, estadísticas, etc, ya que pueden afectar la comprensión del texto por parte de lectores que no estén familiarizados con tales conceptos [21], [22]. Estos trabajos inspiraron una prueba de concepto realizada en este trabajo, detallada en la Sección VII.

En [23] se recopilan muchos de los trabajos en el área de la ST y se puede apreciar que la mayoría se enfoca en el idioma inglés y un aspecto poco explorado es la utilización de corpus que no provengan del género periodístico.

Recientemente se ha estudiado en detalle la performance de los sistemas que han participado en las tareas de identificación de palabras complejas (IPC) que participaron en SemEval, llegando a la conclusión de que los métodos que mejor funcionan son aquellos basados en listas de frecuencias y/o diccionarios y que las anotaciones manuales de los recursos son determinantes para el buen desempeño de los sistemas de IPC o ST [24]. Esto avala y refuerza nuestra decisión de trabajar con listas de frecuencias generadas a partir del corpus que se quiere simplificar.

Cabe destacar también que los trabajos relevados aquí, utilizan métricas para evaluar la facilidad de comprensión de un texto limitándose a usar las que existen para el idioma (principalmente si es inglés) o usan las adaptaciones existentes a otro idioma sin realizar un estudio del impacto que estas puedan tener en el idioma o dominio particular en que serán aplicadas, asumiendo que son apropiadas y coherentes. El objetivo principal de nuestro trabajo es justamente confirmar o descartar esa hipótesis.

Recursos utilizados

En esta sección detallamos los recursos y herramientas utilizados.

A. Corpus

Para el desarrollo de esta investigación se utilizaron extractos de las siguientes obras, las que consideramos tienen distintos niveles de lecturabilidad considerando como audiencia a adultos sin dificultades cognitivas y con un nivel medio de alfabetización:

- El Principito [25], es una novela clásica infantil y fantástica, originalmente escrita en Francés y luego traducida al Español; a pesar de los mensajes abstractos que posee (sobre conceptos como la infancia, la amistad, etc) es conocida popularmente como un cuento para niños por el uso de palabras de poca complejidad en su gran mayoría y estructuras sintácticas sencillas y explícitas. Esta obra consideramos que tiene la mayor lecturabilidad de las cuatro.
- Extractos de la página web del Alto Comisionado de las Naciones Unidas para los Refugiados (ACNUR)[10]. Al igual que otras agencias de la ONU, estos documentos utilizan Español internacional, ya que son creados para el público hispanohablante en general, y tienen características de documentos administrativos legales. Este corpus tiene un índice de lecturabilidad intermedio a bajo.
- Corpus Multilingual UN (Multi-UN): este corpus ha sido compilado y alineado automáticamente para los 6 idiomas oficiales de la ONU y contiene más de 300 millones de palabras por idioma [26].En él abundan nombres propios, definiciones políticas, referencias a legislación, etc, haciendo que resulte en un índice de lecturabilidad intermedio, similar al corpus de ACNUR pero a priori no se puede discernir cuál es de mayor lecturabilidad.
- Corpus paleontología: este texto, generado por el Laboratorio de Paleontología de la Universidad Nacional de Colombia[11] provee información sobre esponjas y cnidarios; para comprenderlo se necesita de un conocimiento especializado en este campo de la ciencia, ya se utilizan términos técnicos, por ejemplo “undulipodios” (prolongaciones citoplasmáticas muy finas hacia el exterior de la célula, y dotadas de movimiento gracias a su composición proteica).

B. Métricas de complejidad

Uno de los pasos previos a la ST es la identificación de palabras o frases complejas, que en la mayoría de los casos se hace de forma manual. Existen algunos índices de complejidad basados en ciertas características del texto y para esta investigación se estudiaron los índices “Spaulding’s Spanish Readability” (SSR) [27], “Lexical Complexity” y “Sentence Complexity Index” (SCI) [28], descriptos a continuación:

Spaulding’s Spanish Readability (SSR)

Esta fórmula mide la complejidad de un texto utilizando el promedio de palabras por línea y el porcentaje de palabras complejas según una lista creada por el autor, la cual contiene palabras que él considera poco frecuentes en el español (y por lo tanto complejas). A mayor complejidad mayor será el valor de retorno de esta formula. El autor propone contar algunas palabras que no se encuentran en la lista como si estuviesen en ella a los fines de simplificar tanto el cómputo del índice como la creación de la lista, y son:

- Nombres de meses y días de la semana.
- Nombres propios apegados a las reglas de los nombre propios para el español.
- Nombres de ubicaciones geográficas

La fórmula es la siguiente:

[10] Agencia de la Organización para las Naciones Unidas (ONU) que trabaja para proteger y asistir a los refugiados en todo el mundo; más detalles en http://www.acnur.org/que-hace/

[11] Disponible en http://www.paleounal.org/doku.php?id=esponjas:esponjas

$$SSR = 1{,}609 \cdot \frac{|w|}{|s|} + 331{,}8 \cdot \frac{|rw|}{|w|} + 22{,}0 \qquad (1)$$

- s es número de oraciones en el texto.
- w es el número de palabras en el texto.
- rw es el número de palabras raras, es decir aquellas que no se encuentran en la lista que oficie como "lista de palabras complejas".

Spaulding acompaña la fórmula con una forma de interpretar los resultados obtenidos, como se muestra en el cuadro 1.

Cuadro 1. Interpretación del puntaje obtenido con SSR

Puntaje SSR	Lecturabilidad
menos de 40	Material muy simplificado
40-60	Muy Fácil
61-80	Fácil
81-100	Dificultad moderada
101-120	Difícil
121 o más	Muy difícil

Lexical Complexity (LC)

Esta métrica también utiliza una lista de palabras para calcular el número de palabras de baja frecuencia, la cual se propone que sea la lista de frecuencia menor a 1000 del corpus Corpus de Referencia del Español Actual (CREA)[12]. Además, toma en cuenta la cantidad y el contenido de las oraciones del texto a medir así como su diversidad léxica. Al igual que en SSR a mayor complejidad mayor será el valor de retorno de esta formula. La fórmula se presenta a continuación:

$$LC = \frac{LDI + ILFW}{2} \qquad (2)$$

$$LDI = \frac{|dcw|}{|s|} \qquad (3)$$

$$ILFW = \frac{|lfw|}{|cw|} \cdot 100 \qquad (4)$$

Donde:

- LDI: Densidad Léxica Indexada

[12] Disponible en http://corpus.rae.es/lfrecuencias.html

- ILFW : Palabras de baja frecuencia indexada
- dcw: Número de diferentes palabras contenidas
- s: Número de oraciones.
- lfw: número de palabras de baja frecuencia (baja frecuencia en CREA ranking menor a 1000)
- cw: Número de tipos de palabras en las oraciones (verbo, sujeto, adjetivos y adverbios).

Sentence Complexity Index (SCI)

Propuesta por Anula (2007) [28], esta métrica mide la complejidad dados el promedio de palabras por oración, y el porcentaje de oraciones marcadas como complejas en el texto indexado. La dificultad al implementar la automatización de esta formula, recae en que cs (número de oraciones complejas) fue creada para catalogar una oración como compleja o no de SSR=1, forma manual. Esto es muy complicado de implementar ya que es un proceso cognitivo muy complejo de simular por medio de la programación, incluso si se contara con un gran corpus del que se pudiera aprender automáticamente.

$$SCI = \frac{ASL + ICS}{2} \qquad (5)$$

$$ASL = \frac{|w|}{|s|} \qquad (6)$$

$$ICS = \frac{|cs|}{|s|} \cdot 100 \qquad (7)$$

- w: Número de palabras.
- s: Número de oraciones.
- cs:Número de oraciones complejas. Esta se calcula manualmente por un apuntador o serie de apuntadores.

C. Herramientas

Para agilizar el estudio de estas métricas es necesario automatizar su cómputo y para ello, es necesario contar con herramientas para el análisis sintáctico y morfológico del español. Luego de evaluar algunas herramientas open source, decidimos que la mejor opción es utilizar Apertium [29], un sistema de traducción automática basado en reglas y originalmente diseñado para lenguajes cercanos, por ejemplo aragonés – catalán, sueco – noruego, español – francés, etc.

Conceptualmente, en nuestro caso podríamos estar traduciendo de “español regular” a “español simplificado”. Sin embargo, en esta etapa exploratoria usamos la versión monolingüe de Apertium para análisis del español[13] ya que ha mostrado mejor performance en términos

[13] Disponible en https://github.com/apertium/apertium-spa/

de cubrimiento léxico, al incorporar un diccionario muy completo, y en términos de análisis morfológico, al contar con mucho conocimiento experto introducido en Apertium por los lingüistas que lo desarrollan.

Una de las principales necesidades para la automatización de métricas es obtener información morfológica de las palabras en las oraciones, para lo cual se utilizó `appertium-spa`, ya que su función `apertium spa-disam-tagger` analiza el texto identificando las palabras con tags, de si la palabra en la oración corresponde a un verbo (`<v>`), sustantivo (`<n>`), adjetivo (`<adj>`), etc.[14]. La figura 1 muestra una comparación de dos versiones de Apertium: la versión como traductor `en-es` y la versión monolingüe (`spa`); se puede ver que `spa` hace un análisis más correcto desde el punto lingüístico, ya que capta contenido de las oraciones (como `<ij>` para interjecciones), mientras que la versión para traducción tiene incorporadas reglas específicas para la traducción automática, por ejemplo, en figura 1 se ve que más se categoriza como pre-adverbios que luego utiliza para posicionar el aumentativo al momento de la generación en inglés, que podría ser “slower”.

Apertium	Resultado: Por favor hable más despacio.
en-es	^Por favor<adv>$ ^hablar<vblex><prs><p3><sg>$ ^más<preadv>$ ^despacio<adv>$
spa	^Por favor<ij>$ ^hablar<vblex><prs><p3><sg>$ ^más<adv>$ ^despacio<adv>$^.<sent>$

Figura 1. Ejemplo de análisis con Apertium como traductor `en-es` y como herramienta monolingüe de análisis del español (`spa`).

La otra herramienta que ha sido utilizada es `PyStemmer`[15], para encontrar la raíz de las palabras antes de su búsqueda en una lista. Esto dado que en los textos se encuentran verbos conjugados y sustantivos en diferentes géneros por ejemplo correr, corrieron, hijos, hijas. Al sacar la raíz de las palabras y buscar si esta se encuentra en el diccionario, facilita el encontrar si una una palabra en cualquier conjugación o género se encuentra en la lista según la necesidad de cada métrica (SSR Spaulding list, LC lfw cuya definición fue mostrada en la parte III-B). Se pudo haber utilizado el lematizador de Apertium pero el enfoque de este no es sacar la raíz de las palabras sino buscar su forma en infinitivo para la traducción durante su pipeline (denominado `mode`), lo cual no resultó muy práctico para nuestro caso.

Metodología

En esta investigación se considera que la simplificación textual está dividida en 4 etapas, como se ilustra en la figura 2 La primera etapa, *Medición de complejidad*, se realiza sobre el texto total (o corpus de entrada), la segunda *Identificación de lineas complejas*, se realiza sobre cada línea del corpus, la tercera es la etapa de simplificación con ayuda de alguna metodología manual o automática y la cuarta, *Verificación de disminución de complejidad respecto al texto original*, es la etapa de evaluación de la efectividad del sistema de ST. En este trabajo abordamos las dos primeras etapas y realizamos una prueba de concepto de la tercer etapa, a los fines de validar los resultados de las etapas anteriores.

[14] La lista completa de símbolos puede consultarse en http://wiki.apertium.org/wiki/List_of_symbols
[15] Disponible en https://pypi.org/project/PyStemmer/

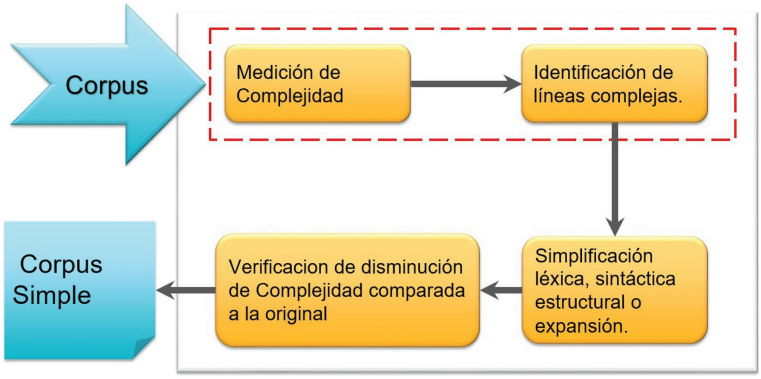

Figura 2. Proceso clásico para ST. Los pasos marcados con línea de puntos son los que se abordan en este trabajo.

Las métricas para medición de índices de complejidad fueron diseñadas originalmente para ejecutarse de forma manual pero con la cantidad de información que quisiéramos procesar no sería práctico realizarlo así. Es por ello, que la primer contribución de nuestro trabajo es la implementación de una solución en *Python* que automatiza las métricas y su aplicación[16]. De acuerdo a las definiciones dadas en la Sección III-B los ingredientes de cada métrica son:

Para SSR:

1. Contar oraciones del texto
2. Contar número, días, mes, lugares geográficos, nombres propios de palabras por oración.
3. Contar palabras raras (que no estén en la lista de Spaulding).

Para LC:

1. Contar número de palabras diferentes en la oración.
2. Contar número de oraciones.
3. Contar número de palabras de baja frecuencia (menor a 1000 en CREA)
4. Contar el contenido de las oraciones (sujetos, adjetivos, verbos y adverbios).

Toda esta información es posible conseguirla aplicando Apertium y PyStemmer al corpus de entrada y basando el análisis morfo-sintáctico en las listas de palabras (de Spaulding o del CREA). Una funcionalidad muy útil de Apertium para la identificación de palabras no tan comunes, es que en la salida puede marcar con *aquellas palabras que no están en su diccionario; si bien esto no garantiza que sean palabras raras o complejas, es un buen indicio dada la amplia cobertura léxica de Apertium.

La solución implementada es open source, toma como entrada un corpus y genera como salida un cuaderno de trabajo con 4 hojas de cálculo: un resumen y una para cada métrica. La primer hoja de resumen contiene tanto el valor global de los índices de lecturabilidad como los valores individuales para cada frase y, de ser pertinente, otras informaciones útiles para el usuario; en esta hoja se muestran los resultados de las primeras n líneas, siendo n ingresada por el usuario al ejecutar el script. La figura 3 muestra un ejemplo de la salida de nuestro sistema en el cual se fijó n = 3. En la segunda, tercera y cuarta hoja se muestran los ranking completo de las métricas SSR, LC y SSCRAW (definida más adelante), respectivamente.

[16] Disponible en https://git.tec.siua.ac.cr/ranaraya/SimplificacionTextualEspanol

	A	B	C
1	Texto Analizado: textoApuntadores		
2	Complejidad LC Total	31.61634137	
3	Complejidad de SSR Total	74.42789474	
4	Top 3 SSR		
5	linea10	175.7113077	
6	linea3	96.06168293	
7	linea9	92.796	
8	Top 3 LC		
9	linea10	69.83333333	
10	linea8	54.5	
11	linea5	53.5	
12	Top 3 SSCRAW		
13	linea10	245.544641	
14	linea8	125.77	
15	linea3	125.5616829	

Figura 3. El resultado de aplicar las métricas a un texto se guarda en una hoja de cálculo conteniendo no sólo los valores sino información adicional que pueda servirle al usuario.

Una vez seleccionados los corpus e implementadas las métricas, la metodología adoptada fue la de medir extractos de los corpus y comparar los resultados iterativamente para comprobar que: las métricas implementadas funcionan debidamente, proveen resultados según la intuición de los investigadores (es decir que son coherentes con la complejidad aparente de los textos seleccionados) y que son capaces de reflejar apropiadamente los cambios producidos luego de operaciones de simplificación (es decir que pueden mantener la coherencia). Estos experimentos y sus resultados se detallan en la siguiente sección.

Experimentos y resultados

A. Estudio de SSR

Para verificar si las métricas reflejan resultados coherentes, es decir alineados con nuestra apreciación de lecturabilidad de cada corpus, se tomó un extracto de 400 palabras de cada uno de los corpus mencionados anteriormente. El cuadro 2 ilustra estos recursos con una oración de cada uno. La elección de estos corpus responde a que quisimos estudiar textos de distinta lecturabilidad considerando como lector final a un adulto con alfabetización media y sin ninguna patología diagnosticada.

Cuadro 2. Ejemplo de los corpus utilizados, que tienen distintos niveles de lecturabilidad.

Principito	Cnidarios	Multi-UN	ACNUR
Enseñé mi obra de arte a las personas mayores y les pregunté si mi dibujo les daba miedo.	Los coanocitos, conocidos también como células de collar, son células globulares provistas de un flagelo o undulipodio (pie oscilante).	Ya se ha establecido la Comisión Mixta de Cesación del Fuego y ya ha empezado a funcionar.	Esta información debe ser lo más detallada posible indicando fechas, lugares, hechos, entre otros.

El resultado de aplicar las métricas a los extractos de 400 palabras se muestra en el cuadro 3. Como se puede ver estos resultados confirman que hay una clara diferencia en la complejidad de los textos afirmando que las métricas brindan un resultado que va en línea con las hipótesis de lecturabilidad que tenemos de los corpus.

Cuadro 3. Cuadro de resultados SSR y LC.

Extracto	Principito	Multi-UN	ACNUR	Esponjas
SSR (Spaulding)	116.79	150.09	195.02	205.60

Un análisis más detallado del índice obtenido para cada corpus muestra que oraciones muy cortas (< 3 palabras) son altamente penalizadas si tienen alguna palabra desconocida según de la lista de Spaulding. Por ejemplo, frases del estilo *Artículo XI, inciso 1* resultan en baja lecturabilidad cuando en realidad es posible que no lo sean. Un caso extremo se muestra en el cuadro 4, donde se detectó que la frase *Ayer me compré un nuevo celular.* obtiene un índice de 86,95 solo por contener la palabra *celular*.

Cuadro 4. Ejemplo de oraciones que se ven penalizadas por ser cortas y tener una palabra que no está en la lista de Spaulding ("celular") versus la misma frase reemplazando la paplabra desconocida por SSR por una conocida

Lista	Texto	Palabras complejas encontradas	SSR	LC
SSR (Spaulding)	Ayer me compré un nuevo celular.	celular	86.95	15.5
SSR (CREA)	Ayer me compré un nuevo carro.	–	31.65	15.5

SSR cataloga la palabra *celular* como compleja, pero llegamos a la conclusión de que esto sucede porque la lista de Spaulding fue creada en 1956, y el lenguaje español ha evolucionado con el paso del tiempo y estos avances lógicamente no son captados por SSR. Esto nos llevó a explorar formas de actualizar esta métrica como se detalla en siguiente sección.

B. Actualización de SSR

Dados los resultados de anteriores se procedió a cambiar la lista que se utiliza en el cómputo de SSR por la lista de frecuencias del CREA completo (descrita en la Sección III), para así enriquecer el diccionario de palabras conocidas y utilizar una versión actual del español. Los resultados de aplicar SSR con esta actualización se muestran en el Cuadro 5; se puede observar que los puntajes bajan mucho pero se mantienen las proporciones sobre la lecturabilidad de los corpus y a pesar del cambio de las listas que son utilizadas como diccionarios, siempre se mantuvo una diferencia marcada entre los documentos. Esto muestra que el cambio propuesto no va en detrimento de la métrica original sino que la adapta al estado actual del lenguaje.

Cuadro 5. Cuadro de resultados SSR: CREA VS. SPAULDING

Extracto	Principito	Multi-UN	ACNUR	Esponjas
SSR (Spaulding)	116.79	150.09	195.02	205.61
SSR (CREA)	45.46	79.82	89.67	101.92

También podemos confirmar nuestra hipótesis de que la lista original ha quedado obsoleta dada la evolución natural del lenguaje, especialmente en dominios tecnológicos y científicos. Se recalculó SSR con la nueva lista para la linea *Ayer me compré mi nuevo celular* y como se muestra en el cuadro 6, ya no hay diferencia entre oraciones que contengan palabras populares como carro o celular.

Cuadro 6. Ejemplo del celular con CREA (Iguala las dos frases)

Texto	Palabras complejas	SSR
Ayer me compré un nuevo celular.	–	31.65
Ayer me compré un nuevo carro.	–	31.65

C. Estudio de LC

Trabajando con las formulas SSR y LC, se pudo notar que SSR al tener una tabla de interpretación de resultados brinda con más claridad una respuesta de complejidad y además brinda mejores resultados al actualizar su diccionario. LC, por el contrario, no cuenta con tabla de interpretación de resultados y no es posible brindar un índice concreto y límite a partir del cual una oración pasa a ser compleja o no. Los resultados de LC se muestran en el cuadro 7. Si comparamos el ordenamiento de los corpus según estos índices de lecturabilidad, vemos que no siempre concuerdan: SSR orden por lecturabilidad descendente Principito < Multi UN < *ACNUR* < *Cnidarios* mientras que LC los ordena *Principito* < *Cnidarios* < *ACNUR* < Multi UN . Esta discrepancia nos llevó a una inspección manual de los extractos para verificar los resultados de LC. Pudimos comprobar que el corpus Multi-UN efectivamente está redactado de forma que contiene estructuras sintácticas más complejas (oraciones largas, con cláusulas subordinadas, muchos acrónimos, etc) que el corpus sobre Cnidarios, que a primera vista parece más complejo (incluso según SSR) por la terminología; sin embargo, como en la mayoría de los textos técnicos altamente especializados, las estructuras sintácticas son básicas y el contenido suele explícito y directo.

Cuadro 7. Resultados de aplicar LC comparados con los resultados mejorados SSR (CREA).

Métrica	Principito	MultiUN	ACNUR	Cnidarios
LC	20.34	35.13	28.76	26.73
SSR (CREA)	45.46	79.82	89.67	101.92

Ya que el objetivo de esta investigación es brindar ayuda en el paso previo a la ST, identificando las lineas más complejas de un texto en forma automática, decidimos explorar formas de fusionar los resultados parciales en uno final, como se describe en la siguiente sección.

D. Estudio intracorpus

Hasta ahora hemos aplicado las métricas para comparar lecturabilidad entre distintos corpus, pero también es necesario estudiar cómo se comportan las métricas entre oraciones de un mismo corpus (intracorpus); dado que se supone más homogeneidad entre ellas, surge la hipótesis de que tal vez no sea tan clara la diferencia entre las oraciones como lo fue entre los corpus.

El interés de un análisis intracorpus reside en que quisiéramos que nuestra solución provea un ranking de oraciones, donde el primer puesto es el de mayor complejidad (menor lecturabilidad) y el último el de menor complejidad, así el usuario podrá decidir qué oraciones simplificar, por ejemplo las top n.

El cuadro 8 muestra tres oraciones del corpus del ACNUR que tomaremos como ejemplo para ilustrar el estudio intracorpus y el cuadro 9 muestra el ranking generado por SSR y LC.

Cuadro 8. Oraciones del corpus del ACNUR que ilustran el estudio de complejidad Intracorpus.

Oración	ID
También quisiera expresar mi reconocimiento a la Misión Africana en Burundi por la destacada labor que lleva a cabo en ese país a pesar de los problemas monumentales que afronta.	A
He pedido al equipo de las Naciones Unidas en Burundi que continúe cooperando estrechamente con mi Representante Especial y coordinando sus actividades con él para ayudar eficazmente al Gobierno Provisional en sus respectivas esferas de competencia.	B
La Misión Permanente de la India ante las Naciones Unidas (Viena) saluda atentamente al Secretario General de las Naciones Unidas y tiene el honor de transmitirle, de conformidad con el artículo IV del Convenio sobre el registro de objetos lanzados al espacio ultraterrestre (resolución 3235 (XXIX) de la Asamblea General, anexo), información relativa al lanzamiento del satélite indio INSAT-3A y la degradación del satélite indio GSLV-D1-GS3 (véase el anexo).	C

Con este ranking, es claro que *C* es la oración de menos lecturabilidad y ambas métricas concuerdan, mientras que la lecturabilidad de la oraciones *A* y *B* difiere entre SSR y LC. Se puede observar que ambas tienen resultados muy cercanos con SSR pero la diferencia es más marcada en LC. Simplemente mirando el largo de *B* podríamos concluir que es más compleja que *A*, sin embargo SSR determinaría lo contrario influenciado.

Cuadro 9. Ranking de lecturabilidad de las oraciones A, B y C de acuerdo a SSR Y LC.

Ranking	SSR(x)CREA	LC(x)
1	142,63 (C)	55,55 (C)
2	81,33 (A)	43 (B)
3	79,92 (B)	29,16 (A)

por la cantidad de palabras desconocidas. Por otro lado, LC es menos sensible a las palabras desconocidas, cambiando de orden el ranking y posicionando *B* como menos comprensible que *A*, más en línea con la apreciación humana.

Teniendo en cuenta estos resultados y considerando que estamos construyendo una herramienta de ayuda a un usuario final deberíamos encontrar una manera de conciliar estos resultados para proveer un único ranking. Además, entendemos que la calidad es un concepto multifacético, por lo que es mejor medir distintos aspectos del mismo objeto para proveer una evaluación integral del mismo. Con esto en mente, evaluamos algunas maneras de combinar SSR y LC, llegando a la conclusión de que una manera sencilla pero eficaz de combinarlas es sumando ambos valores. A esta nueva manera de medir la lecturabilidad la llamamos “Spanish Sentence Complexity Ranking Assistant Work” (SSCRAW), dada por la siguiente ecuación:

$$SSCRAW\ (texto) = SSR(texto) + LC(texto) \quad (8)$$

Aplicando esta nueva métrica, el ranking producido se muestra en el cuadro 10 y se puede notar que el resultado es más objetivo y apegado a la realidad.

Cuadro 10. Ranking aplicando la métrica SSCRAW.

Ranking SSCRAW	SSCRAW(X)
SSCRAW(C)	198,18
SSCRAW(B)	122,92
SSCRAW(A)	110,49

Validación de resultados con anotadores humanos

Luego de verificar la coherencia de LC, SSR y SSCRAW, se realizó un estudio para determinar si efectivamente el resultado de SSCRAW coincide con la percepción de un potencial usuario final de nuestro sistema. Para ello, convocamos a tres revisores independientes (alumnos avanzados de computación del TEC, sin dificultades cognitivas diagnosticadas) para que evaluaran la lecturabilidad del extracto de 400 palabras del corpus Multi-UN utilizado en las secciones anteriores. Se les envió el texto por correo electrónico y les solicitamos que luego de leerlo identificaran las palabras, términos u oraciones que no conocieran o que les resultaran difíciles de comprender. Los resultados de los apuntadores se muestran el Cuadro 11 y la figura 4 muestra la salida del sistema con los resultados globales de SSR, LC y SCRAW, así como el ranking conteniendo las top 3 oraciones más complejas según cada métrica. Aunque no se calculó ningún coeficiente de acuerdo entre evaluadores (como Kappa [30]) se puede notar un alto acuerdo ya que solo los términos de la tabla fueron señalados por los evaluadores y en todos los casos al menos dos de los tres están de acuerdo (*Mayoría*); más aún en el 40 % de los casos el acuerdo es entre los tres jueces (*Unanimidad*).

Cuadro 11. Términos señalados como complejos por los anotadores, algunos por uninamidad y otros por la mayoría de los anotadores. El número entre paréntesis es el número de oración en que se encuentra el término.

Unanimidad (3 anotadores)	*Mayoría* (2 anotadores)
glosario terminológico (4) – PARTEX (8) – moratoria (11) antipersonal (11) – ílicitos (12) – DC21320 (13)	marco regulatorio (2) – UNCCD (4) – marginación (3) – Adición (1) robustecer (6) – servidor (4) –rectificaciones (9) – prorrogó (11) – garantías del Pacto (9)

En las notas entregadas por los anotadores, uno de ellos indicó que la oración 10 le pareció compleja, no por las palabras en forma aislada, sino por la longitud y el estilo (tiempo verbal y entidades nombradas). En ese caso ambos análisis (automático y humano) coinciden en que la línea 10 (reproducida en el cuadro 12) es la más compleja ocupando el primer lugar del ranking.

Cuadro 12. Línea 10 extracto de apuntadore

Línea 10
Cabe señalar que, si el tema de la consolidación de la paz después de los conflictos se examinara con arreglo al formato de evaluación a fondo, serían necesarias evaluaciones separadas de los programas ejecutados por el Departamento de Asuntos Políticos, el Departamento de Operaciones de Mantenimiento de la Paz, la Oficina del Alto Comisionado de las Naciones Unidas para los Refugiados (ACNUR), el Departamento de Asuntos Económicos y Sociales, la Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos y el Departamento de Asuntos de Desarme, entre otros.

Top 3 SSR	
linea10	175.7113077
linea3	96.06168293
linea9	92.796
Top 3 LC	
linea10	69.83333333
linea8	54.5
linea5	53.5
Top 3 SSCRAW	
linea10	245.544641
linea8	125.77
linea3	125.5616829

Figura 4. Resultado de SSR, LC y SCRAW para el extracto analizado por los anotadores humanos. El ranking SCRAW de las 3 oraciones más complejas contiene a la oraciones número 10, 8 y 3

Sin embargo, para el resto de las oraciones no hay acuerdo entre los resultados automáticos y humanos, ya que el ranking de las oraciones más complejas según los anotadores sería oración 10 > oración 11 = oración 4 > oración 9 (calculado contando la cantidad de veces que los anotadores señalaron cada oración como problemática). Esta falta de acuerdo posiblemente se deba a que la tarea no estuvo bien definida para los anotadores y se han enfocado principalmente en las palabras complejas, no tanto en la complejidad global de cada frase. Estos resultados requieren un análisis más profundo y una nueva rueda de evaluación humana.

Prueba de concepto: simplificación de números romanos

Luego de los estudios entre corpus, intra-corpus y de acuerdo entre resultados automáticos y humanos, decidimos realizar una prueba de concepto para ver el impacto en una tarea real de ST.

Entre las estrategias de simplificación están la *expansión*, que consiste en reemplazar contenido comprimido por una versión más descriptiva de tal contenido (por ejemplo reemplazando un acrónimo por las siglas completas) y la *reducción* del contenido problemático, que implica recortar partes del texto a través de distinas técnicas, por ejemplo de resúmenes automáticos. De ellas, la más apropiada para los textos del dominio que consideramos aquí (Multi-UN y ACNUR) es expansión, ya que no deseamos arriesgar pérdida de contenido en los textos.

Durante la revisión de trabajos relacionados (ver Sección II) se observó que la simplificación de expresiones numéricas tiene potencial para aumentar la lecturabilidad. En [21] y [22] proponen modificar las expresiones numéricas reemplazándolas por aproximaciones o expresiones similares aunque no sean exactas, por ejemplo en lugar de dejar *1,9 millones de hogares* lo cambian por *2 millones de casas*. Este tipo de operaciones no son aplicables directamente a nuestro dominio ya que no siempre se puede modificar las cantidades o cifras dado que son datos de tinte legal.

Por lo tanto se procedió a utilizar una estrategia de simplificación de expresiones numéricas sobre aquellos objetos que no den lugar a un cambio de sentido del contenido. En particular se identifican los números romanos y se los convierte en su versión decimal, reemplazando por ejemplo *Artículo X* por *Artículo 10*. La hipótesis subyacente es que esto facilita la comprensión del texto ya que los números romanos no necesariamente son conocidos por algunas personas con bajo nivel de alfabetización.

Para identificar los números romanos utilizamos Apertium, modificándole el diccionario interno y realizando algunas operaciones de post-procesamiento para asegurarnos de que ninguna palabra es confundida con un número romano y, por lo tanto, mal simplificada. El cuadro 13 muestra una frase de ejemplo, donde se toma el texto original del corpus MultiUN, se simplifican los números romanos y se genera un texto simplificado. Se puede observar también en el cuadro 14, que la complejidad se ve disminuida en términos de SCRAW, sugiriendo que estos cambios son positivos y no afectan el contenido de los documentos. Para una confirmación más formal de nuestra hipótesis será necesario realizar una tarea con anotadores humanos.

Cuadro 13. Ejemplo de simplificación de números romanos.

Texto original	Texto simplificado
La Conferencia acoge con beneplácito los resultados de la Conferencia sobre medidas para facilitar la entrada en vigor del Tratado de prohibición completa de los ensayos nucleares (Conferencia del artículo **XIV**), celebrada en septiembre de 2003, y la aprobación de 12 medidas prácticas concretas para promover la pronta entrada en vigor del Tratado.	La Conferencia acoge con bene plácito los resultados de la Conferencia sobre medidas para facilitar la entrada en vigor del Tratado de prohibición completa de los ensayos nucleares (Conferencia del artículo **14**), celebrada en septiembre de 2003, y la aprobación de 12 medidas prácticas concretas para promover la pronta entrada en vigor del Tratado.

Cuadro 14. Revisión de complejidad luego de textos sin simplificar y simplificado

Original	Palabras complejas encontradas	SSCRAW
Simplificado	XIV	113.53
Texto_Simple	–	65.44

Conclusiones y trabajo futuro

En este trabajo abordamos la identificación de palabras y oraciones complejas como paso previo a la ST. Realizamos distintos experimentos para estudiar las métricas *SSR* y *LC* aplicables al español, que son las únicas disponibles dados los pocos recursos existentes para nuestro idioma. Nuestro primer aporte es una herramienta de código abierto y libremente disponible para ayudar a la ST. Esta herramienta permite la automatización de las métricas y genera una planilla de cálculo con los resultados globales más el ranking de las *n* oraciones más complejas para que el usuario decida cuáles simplificar (y en un futuro, posiblemente cómo). Creemos que este aporte es importante dado que, a pesar de haber varios trabajos sobre la ST del español que usan estas métricas, no existe según nuestro conocimiento una herramienta disponible para su cómputo. Para esto hicimos fuerte uso del sistema de traducción automática Apertium en su versión monolingüe para el español, dada su precisión y cubrimiento para el análisis morfo-sintáctio del español, lo cual aceleró considerablemente el trabajo.

Luego de un análisis detallado de los resultados, advertimos que la fórmula SSR se enfoca fuertemente en las palabras desconocidas según su diccionario original, el cual es obsoleto respecto al español actual. Esto motivó la propuesta de una forma de actualizar la *SSR* para que sea más sensible al español contemporáneo cambiando el diccionario por la lista CREA y se obtuvieron mejores resultados.

Al estudiar la métrica *LC* notamos que, al proveer información a nivel morfosintáctico, puede combinarse con *SSR* para dar una medida multifacética de la complejidad. Esta fusión en una nueva métrica la denominamos *SCRAW* y realizamos un pequeño estudio con anotadores humanos para contrastar los resultados automáticos con la mirada del potencial usuario final. Encontramos coincidencias entre las anotaciones humanas y las métricas automáticas para un

caso extremo de complejidad pero no para el resto de los casos anotados. Suponemos que el motivo es la informalidad con que se presentó la tarea a los anotadores y planeamos repetir la evaluación humana en un futuro cercano aplicando los ajustes correspondientes.

Finalmente, realizamos una prueba de concepto en una tarea real de ST para probar la herramienta y la métrica implementadas. Por la naturaleza del corpus utilizado decidimos un enfoque expansionista para el tratamiento de números romanos, los cuales son identificados y reemplazados por números decimales. Los resultados obtenidos son prometedores y constituyen la base para futuras estrategias de simplificación. Como trabajo futuro nos planteamos completar el análisis de *SSR* acompañando el cambio de su diccionario con un estudio de su tabla de interpretación para ver cómo varía, y si fuera necesario actualizarla, estudiar maneras de hacerlo. A pesar de que que las formulas con el uso de la lista CREA tienen un buen rendimiento, otra dirección a explorar en el futuro es la creación de listas especializadas y actualizadas específicamente para cada caso de uso, con el fin de cubrir el vocabulario especializado sobre la respectiva temática y así determinar si las métricas se ajustan mejor a cada dominio.

Referencias

[1] Wikipedia, "Derecho de acceso a la información — wikipedia, la enciclopedia libre," 2018, [Internet; descargado 21-marzo-2018].

[2] G. Paetzold and L. Specia, "Semeval 2016 task 11: Complex word identification," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 560–569.

[3] R. Chandrasekar, C. Doran, and B. Srinivas, "Motivations and methods for text simplification," in *Proceedings of the 16th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1996, pp. 1041–1044.

[4] R. Chandrasekar and B. Srinivas, "Automatic induction of rules for text simplification1," *Knowledge-Based Systems,* vol. 10, no. 3, pp. 183–190, 1997.

[5] M. Dras, "Tree adjoining grammar and the reluctant paraphrasing of text," Ph.D. dissertation, Macquarie University Sydney, 1999.

[6] J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait, "Practical simplification of english newspaper text to assist aphasic readers," in *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 1998, pp. 7–10.

[7] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.

[8] P. T. Quinlan, *The Oxford psycholinguistic database.* University Press, 1992.

[9] W. Hwang, H. Hajishirzi, M. Ostendorf, and W. Wu, "Aligning sentences from standard wikipedia to simple wikipedia," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 2015, pp. 211–217.

[10] Z. Zhu, D. Bernhard, and I. Gurevych, "A monolingual tree-based translation model for sentence simplification," in *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 2010, pp. 1353–1361.

[11] K. Yamada and K. Knight, "A syntax-based statistical translation model," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2001, pp. 523–530.

[12] R. Flesch, "A new readability yardstick." *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[13] K. Woodsend and M. Lapata, "Wikisimple: Automatic simplification of wikipedia articles." in Aaai, 2011.

[14] S. Wubben, A. Van Den Bosch, and E. Krahmer, "Sentence simplification by monolingual machine translation," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1.* Association for Computational Linguistics, 2012, pp. 1015–1024.

[15] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens et al., "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, 2007, pp. 177– 180.

[16] S. M. Aluísio, L. Specia, T. A. Pardo, E. G. Maziero, and R. P. Fortes, "Towards brazilian portuguese automatic text simplification systems," in *Proceedings of the Eighth ACM Symposium on Document Engineering*, ser. DocEng '08. New York, NY, USA: ACM, 2008, pp. 240–248.

[17] H. Saggion, E. Gómez-Martínez, E. Etayo, A. Anula, and L. Bourg, "Text simplification in simplext: Making texts more accessible," *Procesamiento del lenguaje natural*, no. 47, pp. 341–342, 2011.

[18] A. Candido Jr, E. Maziero, C. Gasperin, T. A. Pardo, L. Specia, and M. Aluisio, "Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese," in P*roceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2009, pp. 34–42.

[19] S. Bott, L. Rello, B. Drndarevic, and H. Saggion, "Can spanish be simpler? lexsis: Lexical simplification for spanish," Proceedings of COLING 2012, pp. 357–374, 2012.

[20] B. Drndarevic´, S. Štajner, S. Bott, S. Bautista, and H. Saggion, "Automatic text simplification in spanish: a comparative evaluation of complementing modules," in *International Conference on Intelligent Text Processing and Computational Linguistics.* Springer, 2013, pp. 488– 500.

[21] S. Bautista and H. Saggion, "Can numerical expressions be simpler? implementation and demostration of a numerical simplification system for spanish." in *LREC*, 2014, pp. 956–962.

[22] S. B. Blasco, "Un modelo computacional para la simplificación automática de expresiones numéricas," 2015.

[23] M. Shardlow, "A survey of automated text simplification," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 1, pp. 58–70, 2014.

[24] M. Zampieri, S. Malmasi, G. Paetzold, and L. Specia, "Complex word identification: Challenges in data annotation and system performance," in *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, 2017, pp. 59– 63.

[25] A. Saint-Exupéry, El principito, 2003.

[26] A. Eisele and Y. Chen, "Multiun: A multilingual corpus from united nation documents," in *Proceedings of the Seventh conference on International Language Resources and Evaluation*, D. Tapias, M. Rosner, S. Piperidis, J. Odjik, J. Mariani, B. Maegaard, K. Choukri, and N. C. C. Chair), Eds. European Language Resources Association (ELRA), 5 2010, pp. 2868–2872.

[27] S. Spaulding, "A spanish readability formula," *The Modern Language Journal,* vol. 40, no. 8, pp. 433–441, 1956.

[28] A. Anula, "Tipos de textos, complejidad lingüıstica y facilicitación lectora," in *Actas del Sexto Congreso de Hispanistas de Asia,* 2007, pp. 45–61.

[29] M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. OrtizRojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers, "Apertium: a free/open-source platform for rule-based machine translation," *Machine translation*, vol. 25, no. 2, pp. 127–144, 2011.

[30] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.