LECOLE algorithm: readability, readability, and comprehensibility in Spanish

[Algoritmo LECOLE: legibilidad, lecturabilidad y comprensibilidad en español]

Elena del Pilar Jiménez Pérez – Olivia López Martínez – Javier Corbalán Berná

DOI: 10.18355/XL.2024.17.01.12

Abstract
This article presents the conceptualization of three dimensions surrounding reading: comprehensibility, readability, and legibility. On one hand, readability and legibility for the traceability of reading according to the text, and comprehensibility, on the other hand, as the symbiosis between text and reader. From the perspective of the reader (comprehensibility, ability, and level of a reader to understand a text as it also depends on the skills of the message receiver) and the text itself (legibility and readability or level of difficulty that a text per se entails through its intrinsic formal characteristics) generating two new formulas that define the terms of legibility and readability (specifically in Spanish) since that of comprehensibility exists today and is based on the Common European Framework of Reference for Languages (CEFR), with six levels ranging from A1 to C2 according to TEECLED. Two algorithms that respond to the previous concepts have been developed from the theoretical framework. The suggested formulas of both concepts, readability and comprehensibility, have been validated with texts published in the last 10 years by the Instituto Cervantes, classified by levels A1-C2. The results show that it coincides with the A1-C2 classification of the official DELE exam texts of the Instituto Cervantes in the last 10 years by an average of over 80%.
Key words: Readability, legibility, comprehensibility, algorithm, Spanish

Resumen
Este artículo presenta la conceptualización de tres dimensiones entorno a la lectura: la comprensibilidad, la lecturabilidad y la legibilidad. Por un lado, la lecturabilidad y legibilidad para la trazabilidad de la lectura según el texto, y la comprensibilidad, por el otro, como la simbiosis entre texto y lector. Desde la perspectiva del lector (comprensibilidad, capacidad y nivel de un lector de entender un texto ya que depende también de las habilidades de quien recibe el mensaje) y del texto en sí (legibilidad y lecturabilidad o nivel de dificultad que entraña un texto per se a través de sus características formales intrínsecas) generando dos nuevas fórmulas que acotan los términos de legibilidad y lecturabilidad (específicamente en español) puesto que el de comprensibilidad existe hoy en día y está basada en el Marco Común Europeo de Referencia para las Lenguas (MCER), con seis niveles que van de A1 a C2 según TEECLED. Han sido elaborados desde el marco teórico los dos algoritmos que dan respuesta a los conceptos anteriores. Las fórmulas sugeridas de ambos conceptos, lecturabilidad y comprensibilidad, se han validado con los textos publicados en los 10 últimos años por el Instituto Cervantes clasificados por niveles A1-C2. Los resultados muestran que coincide con la clasificación A1-C2 de los textos de exámenes oficiales DELE del Instituto Cervantes en los 10 últimos años en una media superior al 80%.
Palabras clave: Lecturabilidad, legibilidad, comprensibilidad, algoritmo, español

Introducción

Breve paseo por la historia

Tradicionalmente, los conceptos de legibilidad, lecturabilidad y comprensibilidad de un texto se han mezclado en las fórmulas de lo que en inglés se conoce como readibility, a pesar de existir el término legibility y understandability. No es hasta hace pocos años que se diferencia entre lecturabilidad y comprensibilidad de un texto escrito, aunque no se definan ambos conceptos con precisión con respecto a la legibilidad (Rello, Baeza-Yates et al. 2013). Tampoco se hayan generado fórmulas distintas al respecto con suficiente solvencia (Crossley et. Al., 2008).

Quizás el primer proyecto científico de lecturabilidad lo acometió Thorndike en 1921 sentando las bases teóricas para que, dos años más tarde. Lively and Pressey (1923) publicaran el primer algoritmo para medir dicho concepto. Aunque ya en 1893, Sherman planificó una forma científica y objetiva de enfocar la lectura. En 1928 llega el turno de Vogel y Washburne, quienes definen un método objetivo para graduar el material leíble por los niños. Tres años más tarde, Patty y Painter desarrollan una técnica para medir el vocabulario en libros de texto. Washbourne y Morphett (1938) preceden con su método que incluye una escala de rendimiento lector a la más popular, la fórmula de Flesch, publicada en 1948, precedida a su vez de otras publicaciones (1943, 1945) y que, aun en la actualidad, sigue siendo utilizada con frecuencia incluso en español a pesar de sus múltiples limitaciones (McConnell y Paden, 1983; Power, Sumner y Kearl, 1958). Este mismo año ve la luz la fórmula de lecturabilidad de Dale y Chall, exitosa en el ámbito escolar de la época y algo más precisa que la de Flesch (Power, Sumner y Kearl, 1958) mientras que cuatro años antes Lorge publica su predictor de lecturabilidad. Corre 1952 cuando Klare publica el primero de una serie de artículos sobre la lecturabilidad que, hasta el año 2000, continúan la misma línea. En 1953, Spache genera su fórmula respaldada por el listado de palabras más usadas. Ese mismo año, Taylor defiende el procedimiento Cloze para medir la lecturabilidad, como también Coleman (1975). Tres años más tarde, 1956, Stone publica su crítica a Spache con su medida de dificultad de textos de primaria. En los años 60 destacan Fries, con su Readibility Graph (1963) y The SMOG formula (McLaughling, 1969). Ya en los 70, con The Forcast Formula (Caylor et al. 1973) son las fuerzas armadas norteamericanas las que se interesan por una fórmula para medir la complejidad de sus propios textos.

Desde 1980 se han generado unas doscientas fórmulas y se han publicado más de mil artículos al respecto (DuBay, 2004, 2006), en distintos idiomas como el sueco Lix (Anderson, 1983), no solo en inglés.

ATOS y LEXILE, entre finales de los 90 y principios del presente milenio, son la formulación comercial del concepto; son a la lecturabilidad lo que Goleman al concepto de Inteligencia Emocional (nominado inicialmente por Salovey y Meyer que, a su vez, lo desarrollaron a partir de las Inteligencias Múltiples de Gardner). En 2014 (Milone y Biemiller) se publica la validación comercial de ATOS, atendiendo a dificultades semánticas y sintácticas. Mientras que LEXILE lo publica en 2007, centrándose en las habilidades de comprensión de los lectores, así como en la dificultad sintáctica y semántica del texto, dando como unidad de dificultad el lexile.

En 1952 Gunning publica su The Technique of Clear Writing, con una fórmula para medir la dificultad de lectura de un texto. Lo hizo creando un ejemplo de cien palabras de un texto, dividiendo el número de palabras por el de frases, así como detectando las palabras difíciles (de tres sílabas o más que no fueran nombres propios). Fernández Huerta (1959) traduce la fórmula de Flesch adaptándola al español sin tener en cuenta aspectos básicos como la tipografía y generando controversia al respecto (Law, 2011). Así, siendo L lecturabilidad, P el promedio de

palabras por sílabas y F la medida de palabras por frase, añade las invariables del inglés emitidas por Flesch a su fórmula: L= 206,84 – 0,60 P – 1,02 F. La escala está comprendida entre 0 y 100, siendo inversamente proporcional con la dificultad, a menor valor más difícil. Pero no se puede establecer un paralelismo científico entre fórmulas basadas en lenguas extranjeras, como es el caso del inglés, para medir la lecturabilidad y la comprensibilidad en español, puesto que difieren tanto en forma como en contenido. Además, ya Gwillim (2011) detecta incongruencias en dicha formulación.

Elena Gutiérrez de Polini crea en el 72 probablemente la primera fórmula de lecturabilidad pensada desde el principio de un español para el español. Aunque ella la denomina de comprensibilidad y tiene en cuenta el número de letras, de palabras y de frases en el texto. Aunque ya existieran las aproximaciones de Fry (1968) y Spaulding (1956), revisados por Crawford (1984, 1989) y Gilliam et al. (1980). Es en 1982 cuando se plantea la efectividad de las fórmulas creadas hasta el momento para medir la lecturabilidad en inglés (Davison y Kantor, 1982).

Francisco Szigriszt-Pazos, en su tesis doctoral publicada en 1993, adapta al español la fórmula de Flesch, al igual que hizo Fernández-Huerta (1959), pero la llama perspicuidad (calidad para ser entendido). Inés Barrios (2008), modifica la de Szigriszt-Pazos y crea su Inflesz, por lo que el término de perspicuidad se mantiene. En 2013, McNamara et al. desarrollan una nueva forma de medir la dificultad de un texto midiendo su cohesión. Y Child (2014) publica su web para medir la readability basándose en estudios anteriores como el de Flesch-Kincaid, y concibe la lecturabilidad como algo más práctico que solamente una cuestión teórico-lingüística. Esto se debe a que considera que el nivel de complejidad del texto, su familiaridad, legibilidad y tipografía alimentan su legibilidad. En todo caso, la *readability* se entiende, con sus matices según los autores, como "un atributo del texto escrito" (Begeny y Greene, 2014); o la complejidad de un texto para ser entendido (Campos et al., 2014) y genera multitud de controversias (Benjamin, 2012) al tener en cuenta las habilidades lectoras de las personas que se enfrentan a los textos, ya sean digitales o tradicionales (Zósimo Pena, 2022).

Aunque ninguna de estas definiciones establece una clara frontera entre la legibilidad de un texto, la forma mecánica de lectura como cualidad del propio texto; la lecturabilidad, como el contenido del texto sumado a la tipografía de la lectura, lo mecánico y lo semántico; y la comprensibilidad, entendida como un paso más allá en el que interviene la capacidad del lector de inferir, su comprensión lectora. Y cada vez se complica más la acotación de los términos y su relación, por ejemplo, incluyendo habilidades propias del lector y no del texto, como computar la meoria operativa (Matricciani, 2023).

Legibilidad, lecturabilidad y comprensibilidad

Se ha aceptado que el lenguaje, entendido como sistema de símbolos, consta de dos características fundamentalmente. Por un lado, un componente semántico y por otro, un componente sintáctico (Stenner et al., 2007). Es decir, se estructuran, grosso modo, ruta fonológica, ruta semántica. Por ese motivo, la mayoría de las fórmulas solo tienen en cuenta el vocabulario como componente semántico absoluto y la longitud de la frase como único elemento sintáctico. En este último caso manifestado, en ocasiones, en niveles de lectura. Además, como se ha observado en la introducción, no separan los conceptos de lecturabilidad y legibilidad y comprensibilidad, siendo su uso arbitrario e, incluso, añadiendo otros como perspicuidad (Szigriszt-Pazos, 93), lo que enrarece aún más la delimitación y definición de ambos conceptos. Es necesario realizar una delimitación de qué son y qué relación existe entre ellos dado que

estableciendo un punto de partida en el que se aclare la teoría se puede avanzar hacia la práctica (Jiménez-Pérez, 2022).

Legibilidad

Se puede entender como una característica formal del texto, el nivel de facilidad o dificultad, según se mire, que el escrito muestra de forma homogénea para cualquier lector. Por lo que se puede hablar de un rasgo intrínseco, la idiosincrasia textual en cuanto a la posibilidad de entenderlo objetivamente.

De la legibilidad es responsable el emisor. Así, su estructura de organización, su adecuación, su estilo, su formateo ortotipográfico, su claridad en la exposición, entre otras cuestiones, van a permitir que el nivel de legibilidad sea mayor o menor. Se traduce, además, en unos patrones concretos que se pueden medir. Así, desde un punto de vista algorítmico, la legibilidad se centra en el número de las palabras por frase, la longitud de las sílabas dentro de cada palabra y la acentuación de cada palabra, en primera instancia. Es la medida en que un texto, por sus propios rasgos, puede ser leído por un ser humano o por una inteligencia artificial (en su manifestación de apoyo a las personas, no desde el punto de vista de programación en la lingüística computacional, por ejemplo).

En el ámbito de la prosodia, el español es un idioma eminentemente llano, el francés agudo o el inglés presenta una marcada tendencia al uso de esdrújulas y sobreesdrújulas. La dificultad de las palabras esdrújulas y sobreesdrújulas, se entiende mejor en comparación con el hablante español que se centra en la adquisición del inglés como L2/LE. Por ejemplo, la inusual (para el español) pronunciación sobreesdrújula de *vegetables*, que provoca que en los niveles básicos de aprendizaje se "llanice" por influencia del idioma materno y en ocasiones se tenga que repetir la lectura para una correcta pronunciación. Esto provoca una disrupción en el proceso lector. O la adversidad visual que entraña la acumulación de consonantes como ocurre con rhythms, knightsbridge, catchphrase, etc. El español, en esos términos, es un lenguaje más sencillo; sin ir más lejos, toda acumulación de más de dos consonantes o de uso de agudas, esdrújulas y sobreesdrújulas en exceso, provoca una ralentización del proceso lector, por lo tanto, se puede computar como dificultades a tener en cuenta en una fórmula de legibilidad en español. Así, es más difícil leer palabras españolas como transflor, termorretractil o imperscrutable. Incluso palabras tan usuales como Gabriel producen dificultad en algunos hispanohablantes.

En esta línea, también en las personas con dislexia, se ha demostrado que el uso de palabras más frecuentes y cortas mejoran la capacidad de comprender el texto que leen (Rello, Baeza-Yates, Dempere-Marco, Saggion, 2013).

Mientras que en las fórmulas más complejas publicadas en la actualidad sobre legibilidad se han tenido en cuenta solo ítems como letras o sílabas, nunca su relación, que podría indicar la dificultad de lectura de una sílaba (de una a cinco letras). En ningún caso se ha observado la prosodia teniendo en cuenta, tal como se acaba de mencionar más arriba, que el español es un idioma eminentemente llano, al igual que el francés es agudo o el inglés presenta un número mayor de esdrújulas y sobresdrújulas en su léxico.

Lecturabilidad

Se puede entender la lecturabilidad como una característica de contenido del texto, añadida a la legibilidad del mismo, donde influyen las frecuencias de uso de las palabras, esto es, el conocimiento del vocabulario o, dicho de otro modo, fundamentalmente la riqueza semántica, además de la claridad en la redacción o la complejidad sintáctica.

Por lo tanto, la lecturabilidad va un paso más allá de la legibilidad, procedimiento puramente mecánico, a la que incluye en su propia definición. También es responsible el emisor, sin embargo, en los textos literarios, correría a cargo del editor puesto que debe hacerlo lo más entedible possible según el público, pero atendiendo a cuestiones lingüísticas.

Quizás la pauta más cercana a este concepto la formuló Szigriszt-Pazos en su tesis doctoral sobre la perspicuidad, es decir, la capacidad de estar escrito en un estilo inteligible según el DRAE. Así, la legibilidad sitúa su epicentro en la ruta fonológica (también sintáctica y formal) mientras que la lecturabilidad lo hace en la ruta léxica (semántica).

La confusión entre conceptos como legibilidad, lecturabilidad y comprensibilidad, incluso dificultad textual (Yan, 2023) junto con la falta de acuerdo sobre qué herramientas usar para medirlos, complica el panorama. Por ejemplo, la fórmula de Flesch se sigue usando sistemáticamente para medir estas cualidades en español, incluso en Microsoft Word, a pesar de su rendimiento insatisfactorio en páginas web, como señalan Collins-Thompson y Callan en 2004. Además, la aproximación no holística en la definición de estos términos sugiere, según Begeny y Greene en 2014, que 'la validez de las fórmulas de legibilidad es inconclusa en la literatura científica'. Esto hace imprescindible revisar estos conceptos y considerar de manera inclusiva los avances recientes en la literatura científica sobre el tema. Y la de Matricciani, en ese sentido, tampoco ayuda puesto que complica aún más la realidad añadiendo la memoria operativa sin determinar con exactitud la conceptualización.

Comprensibilidad

La comprensibilidad, sin embargo, hace referencia directa a la comprensión lectora, entendida como la capacidad de un ser humano de entender lo más objetivamente posible lo que un autor ha querido transmitir con su texto (Jiménez-Pérez, 2014). Es decir, lo escrito deja paso al lector y sus habilidades, siendo el punto de inflexión en el que el texto pasa a tomar forma en la realidad lectora de la persona que se enfrenta a su lectura. Así que aquí se tienen en cuenta las habilidades y capacidades del individuo en consonancia con las características del texto; la comprensibilidad es el nexo de unión entre lo mecánico y lo humano.

De esta manera, legibilidad y lecturabilidad se erigen como dos conceptos que se acercan a la comprensión de un texto a través de la comprensibilidad, siendo esta, la antesala de la comprensión y la competencia lectoras (Perfetti y Stafura, 2014; Jiménez-Pérez, 2014), esto es, el nexo de unión en una escala del proceso lector que va desde lo intrínseco al texto hasta lo extrínseco, in crescendo, y desde lo extrínseco del lector hasta lo intrínseco, in decrescendo. De esta forma, se relacionaría la lectura, a través de la competencia lectora, con otras competencias que marca Europa dentro de su marco educativo: la competencia crítica en la lectura (Jiménez-Pérez, 2023; Fernández Millán et al., 2021).

Figura 1: Relación entre conceptos de Legibilidad, Lecturabilidad y Comprensibilidad

Partiendo de lo anterior, cabe destacar que los objetivos del presente estudio son dos: en primer lugar, definir y acotar los conceptos de legibilidad, lecturabilidad y comprensibilidad en español, objetivo que ya ha quedado resuelto en la primera parte de este studio. Y, por otro, validar las fórmulas planteadas para los conceptos de legibilidad y lecturabilidad, puesto que el de comprensibilidad se vincula en última instancia a la comprensión lectora y su fórmula ya está validada (Jiménez-Pérez et al., 2023).

Métodología

El presente estudio es una investigación básica no experimental, es un análisis transversal cualitativo del estado de la cuestión que cerca los conceptos de legibilidad, lecturabilidad para poder establecer un algoritmo que las mida, cuantificando a través de fórmulas sus diferentes niveles en función de las directrices europeas al respecto.
Las fórmulas de legibilidad y lecturabilidad han sido programadas en PHP, que facilita trabajar con base de datos, por ejemplo, las frecuencias de palabras de la RAE. El editor elegido ha sido CoffeeCup Free Editor (porque es gratuito y fácil de usar con resultados óptimos), que muestra una vista previa de lo que se va programando además de ayuda contextual a la escritura del código, siendo una herramienta gratuita de libre acceso relativamente fácil de utilizar. La comprensibilidad, por su parte, viene definida por el concepto de comprensión lectora (Jiménez-Pérez, 2014) y cuantificada en niveles A1-C2 (TEECLED, en prensa).

Herramientas

RAE. Por un lado, la RAE publica en su página web oficial, sección del Corpus CREA (http://corpus.rae.es/lfrecuencias.html), tres listados de uso frecuente de palabras en español, 1.000 palabras, 5.000 palabras y 10.000 palabras, que han sido utilizados en la fórmula de lecturabilidad para el apartado semántico. Dichas secuencias listan según el orden de la frecuencia absoluta y de la frecuencia relativa, y contemplan palabras lexemáticas y no lexemáticas. Se ha respetado la inclusión de palabras no lexemáticas en el apartado semántico por la importancia de las mismas en los textos en vista de que, incluso con la leve alteración del orden, pueden dar significación a una frase u frase. Por ejemplo, no es lo mismo una panda de osos que un oso panda.
PISA. Asimismo, los textos seleccionados se han organizado según las últimas directrices PISA (2015) para la clasificación de los textos según el tipo: narrativo, descriptivo, expositivo, argumentativo, instructivo y con diálogo; y según los géneros mediático, literario, de comunicación interpersonal, institucional, laboral y académico. Se ha guardado equilibrio entre los grupos de textos, siendo el mínimo 83 (textos laborales) y 91 el máximo (literarios). Según el tipo: narrativo (86), descriptivo (85), expositivo (86), argumentativo (86), instructivo (84), con diálogo (84); según el género: mediático (84), literario (91), de comunicación interpersonal (88), institucional (84), laboral (83) y académico (84). La división entre tipo y género hace que los textos sean distintos entre sí, esto es, los textos computados dentro el apartado literario no se contabilizan dentro de narrativos, por ejemplo. Se han tenido en cuenta desde recetas, instrucciones, mapas, anuncios, entradas de redes sociales o mensajería hasta literatura, periodismo, publicidad, ciencia y tecnología o religión.

La herramienta resultante de la programación de las fórmulas de legibilidad y lecturabilidad ha sido colgada en la siguiente dirección: comprensionlectora.es/leleco (antes en andaluces.es/lectucompre/indice.php).

Procedimiento

Se han tenido en cuenta una serie de conceptos importantes para la formulación tanto de la legibilidad como de la lecturabilidad, que se detallan a continuación en función de la dificultad en paradigma *bottom up*: dificultad silábica, dificultad prosódica, dificultad léxica, dificultad sintáctica, dificultad semántica.

Creación de los dos algoritmos

Una palabra compuesta por dos sílabas simples, consonante + vocal ("mamá"), es más fácil de pronunciar y, por lo tanto, de leer, que una palabra con 5 sílabas y estructura de consonante + vocal + x consonantes ("inescrutable"); por lo tanto, la dificultad silábica se considera un ítem fundamental en la formulación de la legibilidad.
Igualmente, la pronunciación de las palabras, básicamente en su tonicidad, influyen en la facilidad o dificultad a la hora de leer un texto, en su legibilidad. Tomando como referencia los listados RAE de frecuencia de palabras en español utilizados para validar la herramienta, el porcentaje de palabras llanas en un texto medio oscila entre un el 64.9% para la base de datos de 1.000, 69.5% para 5.000 y 70.2% para 10.000, por lo tanto, la dificultad prosódica desciende a mayor número de palabras. Para los niveles MCER se hace media aritmética entre el primer valor y el segundo (64.9-69.5) para los tres primeros niveles (A1/A2/B1) y de igual forma entre el segundo y tercer valor (69.5-70.2) para los tres restantes (B2/C1/C2).

La dificultad silábica, oscila desde "Mi mamá me mima" de Micho, cuya dificultad silábica es de 1.9 hasta un artículo científico médico con 2.5, el rango oscila entre los 2.2 del Ratón Pérez, los 2.1-2.2 de entradas de Twitter y Facebook respectivamente, los 2.4 de Quevedo o Góngora, o los 2.3 y 2.4 de La Vanguardia y El Mundo respectivamente. Así, atendiendo a los niveles de MCER, las correspondencias serían A1: hasta 2, A2: 2.01 hasta 2.1, B1: 2.101 hasta 2.2, B2: 2.201 hasta 2.3, C1: 2.301 hasta 2.4 y C2: más de 2.4.

La dificultad léxica se extrae como una invariable dependiente de cuatro niveles establecidos a partir de la media de las frecuencias publicadas por la RAE resultantes del listado citado anteriormente. Tramos 1) hasta 1.000; 2) de 1.000 a 5.000; 3) de 5.000 a 10.000, 4) que contemplen a partir de un 10% no reconocidas en esas primeras 10.000.
En el apartado de dificultad sintáctica se ha obtenido el ítem realizando el promedio de palabras de los 126 textos con una media de 653 palabras/texto. En textos de extensión larga se ha seleccionado las 1.000 primeras palabras (textos literarios o académicos) y se ha computado el total de las palabras en los que no alcanzaban ese dato (por ejemplo, de comunicación interpersonal: como un correo electrónico, o mediático: como un tweet/WhatsApp, etc.). La media de palabras por frase de este grupo de textos ascienda a 23.5, siendo el umbral mínimo las 4 de "Mi mamá me mima "y 42 La Vanguardia. Fluctúan los textos de referencia Quevedo 28, Góngora 31 o 22 Facebook.

Por último, la dificultad semántica se obtiene directamente de las frecuencias ya citadas de la RAE. A1: menos de las 1.000, A2: 1.000-3.000, B1:3.000-5.000, B2: 5.000-7.500, C1: 7.500-10.000 y C2: 10.000 más un 10% no incluidas en los listados

anteriores. Las secuencias 3.000 y 7.500 se obtienen de media aritmética entre 1.000 y 5.000, 5.000 y 10.000. Sthal (2003) ya señala que contar con palabras fáciles o difíciles solo según su longitud, por sí solas, no determinan la dificultad de un texto.
La RAE publica tres listados de uso frecuente de palabras en español, 1.000 (nivel 1), 5.000 (nivel 2= 5.000-1.000) y 10.000 palabras (nivel 3=10.000-4.000), (nivel 4= nivel 3 + hasta 10% no incluidas en las 6.000).

De esta forma, la composición de los ítems dentro de la fórmula responde a los siguientes niveles:

1. Dificultad silábica (DSL), que se obtendría dividiendo el número de letras por el número de sílabas, respectivamente. Así, atendiendo a los niveles de MCER para la comprensión de textos (TEECLED, en prensa) las correspondencias serían A1: hasta 2,000, A2: >2 - <=2.1, B1: >2.1 - <=2.2, B2: >2.2 - <=2.3, C1: >2.3 - <=2.4 y C2: >2.4.
2. Dificultad prosódica (DP), que se obtendría según el porcentaje de palabras llanas. A1: -64.9%, A2: 64.9%-67.2%, B1: 67.2%-69.5%, B2:69.5%-69.8%, C1: 69.8%-70.2%, C2: >=70.2%.
3. Dificultad léxica (DL), que se obtendría dividiendo el número de sílabas por el número de palabras, siendo la media comparable la obtenida en las 1.000, 5.000 y 10.000 palabras publicadas en la RAE, donde los tramos contemplan 1.000-5.000 y 10.000-5.000, además de los umbrales superior e inferior. A1: -2.44, A2: 2.44-2.94, B1: 2.94-3.06, B2: 3.06-3.09, C1: 3.09-3.25, C2: >= 3.25.
4. Dificultad sintáctica (DST), que se obtendría dividiendo el número de palabras por el número de oraciones. Según nomenclatura MCER, A1: <4, A2: >=4 - <10, B1: >=10 - <20, B2: >=20 - <30, C1: >=30 - <42, C2: >=42.
5. Dificultad semántica (DSM), que se corresponden con los niveles de frecuencias publicados por la RAE, a saber, la RAE publica tres listados de uso frecuente de palabras en español, Según los niveles MCER, A1: menos de las 1.000, A2: 1.000-3.000, B1:3.000-5.000, B2: 5.000-7.500, C1: 7.500-10.000 y C2: C1 más un 10% no incluidas en los listados anteriores.

En la fórmula de legibilidad se ha tenido en cuenta la dificultad silábica y la pronunciación entendida como dificultad prosódica en relación con la dificultad léxica y semántica. El resultado de la suma de los niveles obtenidos según MCER en DSL, DL, DP, DS y DSM se ponderan siendo los valores enumerativos correlativos. A1: 1, A2: 2, B1: 3, B2: 4, C1: 5, C2: 6; cuya suma se divide entre los cinco ítems (legibilidad) o 6 (lecturabilidad). Así, el resultado correspondiente a un B1, B2, B1, B2, B1 (que se obtienen en las diferentes dificultades recogidas en el algoritmo) para legibilidad corresponde a un B1: 3.4.

FÓRMULA LEGIBILIDAD	FÓRMULA LECTURABILIDAD
Niveles DSL +DP+DL/ 3=LG	LG+DST+DSM/ 3= LC

Validez comparativa

Para la comprobación de la validez de la formulación de ambos algoritmos, legibilidad y comprensibilidad, se utilizaron los textos publicados por el Instituto Cervantes (IC) en sus exámenes oficiales de los últimos diez años, que se encuentran publicados en su espacio web: https://examenes.cervantes.es/es/dele/preparar-prueba. Dichas pruebas se ajustan a las directrices que la Unión Europea publica al respeto de la enseñanza de las lenguas en su Marco Común en https://www.coe.int/en/web/language-policy/home. Estos textos se encuentran

categorizados en los niveles anteriormente citados: A1, A2, B1, B2, C1, C2 por lo que la comprobación se efectúa de forma paralela e inequívocamente las correspondencias son pertinentes.
Puesto que la clasificación del Instituto Cervantes no atiende a las características que se plantean en el presente estudio de dificultades prosódicas, silábicas, léxicas, semánticas y sintácticas, se procede a comprobar el resultado final de las fórmulas de legibilidad y lecturabilidad, con respecto a la clasificación única y global que el Instituto Cervantes realiza de cada texto al incluirlos como forma de evaluación en un nivel determinado dentro del arco A1-C2. Cabe destacar que el Instituto Cervantes se inclina en los aspectos lectores a simplificar los niveles a 3: inicial para A (incluye A1 y A2), intermedio para B (B1 y B2) y avanzado para C (C1 y C2).

Resultados

		Instituto Cervantes					
		INCIAL		INTERMEDIO		AVANZADO	
		A1	A2	B1	B2	C1	C2
L	A	78%/75%					
E			82%/81%				
C	B			90%/83%			
O					79%/82%		
L	C					88%/77%	
E							71%/85%

Tabla 1. Interacción niveles textos IC y LECOLE dividida en LG y LC

Se ha medido en porcentajes la coincidencia de Legibilidad (LG) y Lecturabilidad (LC) debido al bajo número de textos publicados en los últimos 10 años en la web del Instituto Cervantes. Para realizar la medición, se ha programado las formulas en PHP y se han introducido los textos ya clasificados según directrices del IC. Se han redondeado los porcentajes para evitar decimales, inferior a .5 se redondea a la baja, igual o superior al alza.
En la comprobación por niveles inicial (A), intermedio (B) y avanzado (C) del análisis comparativo de los textos del Instituto Cervantes y los de los algoritmos, que incluyen su división en A1-A2, B1-B2, C1-C2, se observa una coincidencia media del 81.33%.
LG. En el nivel inicial, la coincidencia con la clasificación del IC es del 80%, siendo más baja en el nivel A1 (78%) que en el A2 (82%). Sin embargo, dicho porcentaje se eleva hasta el 85% en el nivel intermedio, siendo el B1 (90%) el nivel que más coincide de todos los niveles, el B2 se mantiene en el 79%. El último nivel, el avanzado, muestra una similitud media del 80%, donde el C1 arroja mejores resultados (88%) y el C2 un 71%.
LC. En esta dimensión, la coincidencia con la clasificación del IC varía levemente de la LG. Así, en el nivel inicial, la media es del 81% en A, siendo del 75% en A1 y del 81% en A2. Mientras que en el nivel intermedio suben los porcentajes, una media del %, donde B1 coincide en el 83% y B2 en el 82%. Llegados al nivel 3, la media es de 81%, repartida entre C1 con el 77% y C2 con el 85%.

Conclusión

Al igual que en la comunicación, debe haber un consenso en el código entre emisor y receptor de un mensaje, en el ámbito científico debe existir ese mismo acuerdo en los puntos de partida, las bases, que son los conceptos. A día de hoy, se han venido

utilizando en español los términos de legibilidad, lecturabilidad y comprensibilidad en terreno científico con arbitrariedad. El primer objetivo de este trabajo era establecer la definición, relación y limitación entre sí de los conceptos legibilidad, lecturabilidad y comprensibilidad. A partir de la definición teórica y la contextualización, se pasa al siguiente nivel, la validación de las fórmulas que dan enfoque práctico a dichas definiciones.

Este estudio surge de la necesidad de establecer un punto de partida útil para que editoriales, profesorado y alumnado puedan avanzar en el proceso lector (Muñoz, 2006) a partir del consenso mediante una herramienta que se adapta a los niveles establecidos en la Unión Europea, algo aún no realizado por ninguna otra. Una necesidad altamente demandada por el profesorado (Meyer, 2003). Aunque la intención es que sirva de base no solo en el espacio educativo sino en el día a día de aquellos lectores que necesiten conocer de antemano la dificultad del texto al que se enfrentan.

Como limitación, es necesario tener en cuenta que el nivel de dificultad puede variar de nivel en un mismo libro dependiendo del fragmento que se elija (McConnell y Paden, 1983) ante la imposibilidad real de que un emisor o escritor de un texto largo mantenga siempre el mismo nivel. Por ese motivo, se hace necesario escoger tres fragmentos en el caso del análisis de novelas, por ejemplo, del principio, de la mitad y del final. Esta investigación se ha visto beneficiada por la breve extensión de los exámenes del IC: se han usado los textos completos al no ser excesivamente largos (existe una limitación de 1.500 palabras, para el uso práctico de textos largos 500 iniciales, intermedios y finales). Además, se ha de tener en cuenta que no siempre el emisor del texto consigue transmitir con rigurosidad lo que objetivamente desea expresar (Bruce, Rubin & Starr, 1981), por lo que controlar esa parcela de la realidad es complicado.

En futuras investigaciones y con la ayuda del Big Data, así como de la minería de datos y los ordenadores cuánticos se podría afinar aún más en la validación del algoritmo de las dos fórmulas, puesto que se tendría acceso a los matices con mucha más profundidad. Además, establecer unas equivalencias de niveles en textos y lectura facilitará la integración de textos escolares según el nivel de su alumnado, lo que ha sido de vital importancia desde siempre en el ámbito de la enseñanza (Cline, 1972).

Agradecimientos

Esta investigación se ha realizado al amparo de la Asociación Española de Comprensión ectora, por la IP del equipo de investigación Hum-1048, durante la estancia de investigación para la tesis doctoral de la profesora Jiménez-Pérez en la UM.

Bibliographic references

Anderson, J. (1983). Lix and rix: Variations on a little-known readability index. Journal of Reading, 26(6), 993-1022. https://doi.org/10.1111/j.1467-9817.1983.tb00238.x

Begeny, J., & Greene, D. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? Psychology in the Schools, 51(2), 198-215. https://doi.org/10.1002/pits.21740

Benjamin, R. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24(1), 63-88. https://doi.org/10.1007/s10648-011-9181-8

Barrios, I. (2008). Validación de la Escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes. An Sist Sanit Navar 31(2), 135-152.

Bruce, B.C., Rubin, A. & Starr, K.R. (1981). Why readability formulas fail. IEEE Transactions on Professional Communication, PC-24, 50-52. https://doi.org/10.1109/tpc.1981.6447826
Campos, D., Contreras, P., Riffo, B., Véliz, M. & Reyes, A. (2014). Complejidad textual, lecturabilidad y rendimiento lector en una prueba de comprensión en escolares 99 adolescentes. Universitas Psychologica, 13 (3), 1135-1146. http://dx.doi.org/10.11144/Javeriana.UPSY13-3.ctlr
Chall, J. (1995). Readability revisited: The new Dale-Chall readability formula. Brookline Books/Lumen Edition.
Child, D. (2014). Readability-Score.com. https://readability-score.com/
Cline, T.A. (1972). Readability of Community College Textbooks. Journal of Reading 16(1), 33-37. https://doi.org/10.1080/10862967209547033
Coleman, M. (1975). A computer readability formula designed for machine scoring. Applied Psychology, 60(2), 283-284. https://doi.org/10.1037/h0076540
Collins-Thompson, K., & Callan, J. (2004). A language modeling approach to predicting reading difficulty. Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Boston, USA. 193-200
Crawford, A. (1984). A Spanish language Fry-Type readability procedure. Elemantary level. Bilingual Education Paper Series, 7. https://files.eric.ed.gov/fulltext/ED273119.pdf
Crawford, A. N. (1989). Fórmula y gráfico para determinar la comprensibilidad de textos de nivel primario en castellano. https://www.researchgate.net/publication/268420792_Formula_y_grafico_para_determinar_la_comprensibilidad_de_textos_del_nivel_primario_en_castellano
Crossley, S.A., J. Greenfield, and McNamara, D.S. (2008). Assessing Text Readability Using Cognitively Based Indices. Tesol Quarterly. 42(3), 475-493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x
Dale, E., Chall, J. S. (1948). A formula for predicting readibility. Educational Research Bulletin, 2(1), 20-28.
Dale, E., Chall, J. S. (1948). A formula for predicting readibility: instructions. Educational Research Bulletin, 2(2), 37-54.
Davison, A., y Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. Reading Research Quarterly, 17(2), 187-209. https://doi.org/10.2307/747483
DuBay, W. H. (2004). The principles of Readibility. Impact-Information: California.
DuBay, W. H. (2006). The classic Readibility studies. Impact-Information: California.
Fernández Millán, G., García Guirao, P., & López Martínez, O. (2021). El pensamiento crítico aplicado a "El Lazarillo de Tormes a través del debate en 3º de la ESO. Investigaciones Sobre Lectura,16, 32-50. https://doi.org/10.24310/isl.vi16.12870
Milone, M. & Biemiller, A. (2014). Development of the ATOS® Readability Formula. Reinassence.
Fernández-Huerta J. (1959). Medidas sencillas de lecturabilidad. Consigna, 214, 29-32.
Flesch, R. (1949). The art of readable writing. Harper.
Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 3, 221-233. https://doi.org/10.1037/h0057532
Flesch, R. (1945). How copy writers can use readability tests. Nueva York: Autor.
Flesch, R. (1943). Marks of readable style: A study in adult education. Columbia University: Teachers College.
Fry, E. (1968). A Readability formula that saves time. Journal of Reading, 11, 513-578.

Gilliam, B., Peña, S., & Mountain, L. (1980). The Fry Graph Applied to Spanish Readability. The Reading Teacher, 33(4), 426-430.
Gunning, R. (1952). The Technique of Clear Writing. McGraw-Hill. https://archive.org/details/techniqueofclear00gunn/mode/2up
Gutiérrez de Polini, L.E. (1972). Investigación sobre lectura en Venezuela. Ministerio de Educación, Caracas.
Jiménez-Pérez, E. (2014). Comprensión lectora VS Competencia lectora: qué son y qué relación existe entre ellas. Investigaciones Sobre Lectura, 1, 64-75. https://doi.org/10.24310/revistaisl.vi1.10943
Jiménez-Pérez, E. (2023). Pensamiento crítico y competencia crítica. Investigaciones Sobre Lectura, 18(1), 1-26. https://doi.org/10.24310/isl.vi18.15839
Jiménez-Pérez, E. (2024). TEECLED. Investigaciones Sobre Lectura, en prensa.
Klare, G. R. (1952). Measures of the readability of written communication: An evaluation. The journal of educational Psychology, (43)7, 385-399. https://doi.org/10.1037/h0058972
Law, G. (2011). Error in the Fernandez Huerta Readability Formula. https://linguistlist.org/issues/22/22-2332.html
Lively, B. A. & Pressey, S.L. (1923). A method for measuring the vocabulary burden en textbooks. Educational Administration and Supervisión, 9, 389-398.
Liu, Y. (2023). Readability and adaptation of children's literature: an interpersonal metaphor perspective. Journal of World Languages, 28, 3-22. https://doi.org/10.1515/jwl-2022-0039.
López Pena, Z. (2022). Una propuesta multimodal para la lectura de textos digitales en el contexto de la asignatura Lengua Castellana y Literatura en ESO. Investigaciones Sobre Lectura, 1(17), 21-39. https://doi.org/10.24310/isl.vi17.14475
Lorge, I. (1944). Predicting readibility. Teachers College Record, 45, 404-419.
Matricciani, E. (2023). Readability Indices Do Not Say It All on a Text Readability. Analytics, 2, 296-314. https://doi.org/10.3390
analytics2020016
McLaughlin, G. (1969). SMOG grading - A new readability formula. Reading, 12(8), 639-646.
McNamara, D.S., et al. (2013). Coh-Metrix, 3.0. http://www.cohmetrix.com.
McConnell, C. and D.W. Paden (1983) Readability: Blind Faith in Numbers? Journal of Economic Education, 14(1), 65-71. https://doi.org/10.1080/00220485.1983.10845007
Meyer, B.J. (2003). Text coherence and readability. Topics in Language Disorders, 23(3), 204-224. https://doi.org/10.1097/00011363-200307000-00007
Muñoz, M. (2006). Legibilidad y variabilidad de los textos. Boletín de Investigación Educacional, Pontificia Universidad Católica de Chile, 21 (2) 13-26.
Muñoz, M. y Muñoz, J. (2006). Legibilidad Mμ. Viña del Mar, Chile.
Patty, W. W., & Painter, W. I. (1931). A technique for measuring vocabulary burden in textbooks. Elementary School Journal, 28, 373-381.
Perfetti, C. & Stafura, J. (2014). Word knowledge in a theory of Reading comprehension. Scientific Studies of Reading, 18(1), 22-37. https://doi.org/10.1080/10888438.2013.827687
Power, R., Sumner, W., & Kearl, B. (1958). A Recalculation of Four Adult Readability Formulas. Journal of Educational Psychology, 49(2), 99-105.
Real Academia Española, (2020). "Banco de datos (CREA)," Real Academia Española, http://www.rae.es.
Rello L., Baeza-Yates R., Dempere-Marco L., & Saggion H. (2013) Frequent Words Improve Readability and Short Words Improve Understandability for People with Dyslexia. In: Kotzé P., Marsden G., Lindgaard G., Wesson J., Winckler M. (eds),

Human-Computer Interaction - INTERACT 2013. INTERACT 2013. Lecture Notes in Computer Science, 8120. Springer, Berlin, Heidelberg
Stahl, S. A. (2003). Vocabulary and readability: How knowing word meanings affects comprehension. Topics in Language Disorders, 23(3), 241-247. https://doi.org/10.1097/00011363-200307000-00009
Sherman, I. A. (1893). A manual for the objetive study of English prose and poetry. Ginn and Company: New York.
Spache, G. (1953). A New Readability Formula for Primary-Grade Reading Materials. The Elementary School Journal, 53 (7), 410-13. https://doi.org/10.1086/458513
Spaulding, S. (1956), A Spanish Readability Formula. The Modern Language Journal, 40: 433-441. https://doi.org/10.1111/j.1540-4781.1956.tb02145.x
Stenner, A. J., I. Horabin, D. R. Smith, & Smith, R. (2007). The lexile framework for Reading. Theoretical Framework and Development. Durham, NC: Metametrics, Inc
Stone, C. R. (1956). Measuring Difficulty of Primary Reading Material: A Constructive Criticism of Spache's Measure. The Elementary School Journal, (57) 1, 36-41. https://doi.org/10.1086/459497
Szigriszt-Pazos, F. (1993). Sistemas predictivos de legibilidad del mensaje escrito: fórmula de perspicuidad. Tesis doctoral. UCM: Madrid.
Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 30, 415-433. https://doi.org/10.1177/107769905303000401
Thorndike, E. L. (1921). The teacher Word Book. Columbia University.

Words: 6492
Characters: 42 831 (24 standard pages)

Corresponding author
Elena del Pilar Jiménez Pérez
Didáctica de las Lenguas, las Artes y el Deporte
Facultad de CC.EE.
Universidad de Málaga
Campus de Teatinos s/n
29071 Málaga
Spain

Olivia López Martínez
Universidad de Murcia
Murcia
Spain
https://orcid.org/0000-0002-9819-8005

Javier Corbalán Berná
Universidad de Murcia
Murcia
Spain
https://orcid.org/0000-0001-6614-0522