UNIVERSITAT POLITÈCNICA DE VALÈNCIA

_ TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE UPV INGENIEROS DE TELECOMUNICACIÓN

ANÁLISIS DE ALGORITMOS PARA DETERMINAR EL NIVEL DE COMPLEJIDAD DE TEXTOS SANITARIOS Y RECOMENDACIONES PARA MEJORAR EL EMPODERAMIENTO DE UN PACIENTE

Autor: Alejandro Gallego Andrés

Tutor: Vicente Traver Salcedo

Cotutor: Manuel Traver Salcedo

Trabajo Fin de Grado presentado en la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universitat Politècnica de València, para la obtención del Título de Graduado en Ingeniería de Tecnologías y Servicios de Telecomunicación

Curso 2016-17

Valencia, 12 de septiembre de 2017

Resumen

El objetivo del presente estudio es el desarrollo de un programa capaz de determinar la complejidad de un texto sanitario y la evaluación de los resultados proporcionados por buscadores de Internet ante consultas de información médica utilizando dicha herramienta. Primero se realizó una investigación del estado del arte, determinando las fórmulas para el cálculo de la legibilidad existentes y las herramientas disponibles que hacen uso de estas fórmulas, y se elaboró una tabla comparativa de sus características. Posteriormente se probaron estas herramientas sobre un conjunto de muestras con un nivel de complejidad ya asignado por un profesional de la salud, para determinar que algoritmo obtenía los resultados más acertados. Seguidamente se desarrolló un programa en Python que implementa dicho algoritmo. El programa recibe una URL a una web y obtiene la legibilidad lingüística del contenido, guardando la información en un fichero XML. El siguiente paso fue obtener información sobre las palabras clave de información médica que más buscan los pacientes en Google, y analizar una muestra de artículos con la herramienta que se ha desarrollado. Se concluyó que, en media el 41.43% son fáciles, el 30% son de dificultad media y el 28.57% son difíciles.

Resum

L'objectiu del present estudi és l'elaboració d'un programa capaç de determinar la complexitat d'un text sanitari i l'avaluació dels resultats proporcionats per cercadors d'Internet al realitzar consultes d'informació mèdica utilitzant aquesta eina. Primer es va realitzar una recerca de l'estat de l'art, determinant les fórmules per al càlcul de la llegibilitat existents i les eines disponibles que fan ús d'aquestes fórmules, i es va elaborar una taula comparativa de les seues característiques. Posteriorment es van provar aquestes eines sobre un conjunt de mostres amb un nivell de complexitat ja assignat per un professional de la salut, per tal de determinar que algorisme obtenia els resultats més encertats. Seguidament es va elaborar un programa en Python que implementa aquest algorisme. El programa rep una URL a una web i obté la llegibilitat lingüística del contingut, guardant la informació en un fitxer XML. El següent pas va ser obtenir informació sobre les paraules clau d'informació mèdica que més cerquen els pacients a Google, i analitzar una mostra d'articles amb l'eina que s'ha programat. Es va concloure que, en mitjana el 41.43% són fàcils, el 30% són de dificultat mitjana i el 28.57% són difícils.

Abstract

The objective of this study is the development of a software capable of determining the readability of a health text and the evaluation of the results provided by Internet search engines. First, an investigation of the state of the art was carried out, determining the readability formulas and the available tools that make use of these formulas, and a comparative table of their characteristics was elaborated. These tools were then tested on a set of samples with a level of complexity assigned by a health professional, to know which algorithm obtains the most successful results. Then, a Python program has been developed that implements this algorithm. The program receives a URL to a web and obtains the readability of the content, saving the information in an XML file. The next step was to get information about which are the medical information keywords that most patients are looking for in Google, and to analyze a sample of articles with the tool that has been developed. It was concluded that on average 41.43% of the search results are easy, 30% have a medium difficulty level and 28.57% are hard to understand.

Índice

Índice de figuras:

Índice de tablas:

1. Introducción:

1.1. Información sanitaria y empoderamiento del ciudadano

Actualmente son cada vez más los pacientes que hacen uso de Internet como fuente de información médica[1], esto le permite conocer mejor las patologías existentes, los tratamientos, las medidas de protección frente a éstas y demás información que puede resultar importante para él/ella, y facilitar la comunicación entre profesionales y pacientes.

La tendencia en los últimos años consiste en involucrar cada vez más al paciente en la toma de decisiones con lo que respecta a aspectos de la salud de éste, produciéndose el nacimiento de la figura del paciente empoderado. "Un paciente empoderado es un paciente con capacidad para decidir, satisfacer necesidades y resolver problemas, con pensamiento crítico y control sobre su vida. Y todo ello se consigue, en primer lugar, con el conocimiento"[2]. El paciente empoderado debe recibir información de una fuente de confianza y de su comprensión, ya que, en el caso contrario, la disposición del paciente de gran cantidad de información podría resultar contraproducente y afectar negativamente a la relación entre médico y paciente. Actualmente es común entre los profesionales de la salud la sensación de deshumanización de su persona e incluso desconfianza por parte del paciente, al disponer éste de información de Internet que puede no coincidir con la del médico. (R & Luz, 2005)

Por estas razones, la información que recibe el paciente ha de ser siempre veraz y contrastada por profesionales de la salud, sin embargo, el usuario de Internet medio, consulta información haciendo uso de buscadores no especializados en medicina, obteniendo información de fuentes que podrían no ser confiables o estar destinadas, por su complejidad, a profesionales de la salud. Además, en la red se encuentran a menudo artículos con información no contrastada o con datos erróneos que no han sido redactados por personas cualificadas y que pueden confundir al paciente y/o inducirle a error en la toma de decisiones colaborativa con su médico[3].

En este aspecto, los médicos y profesionales del sector trabajan para ofrecer una información accesible y siempre objetiva a sus pacientes (Barca Fernández et al., s. f.). De esta necesidad nacen proyectos encargados de recopilar artículos e información de interés que son supervisados por profesionales acreditados en el campo de la salud.

Salupedia es un proyecto colaborativo con el objetivo de crear un lugar donde profesionales (médicos, enfermeros, psicólogos, etc.) recomienden contenidos, ya existentes en la red, a pacientes, familiares y ciudadanos en general, asegurando en la medida de lo posible que se trata de información fiable[4].

La web de Salupedia dispone de una gran cantidad de artículos de fuentes diversas y cuyo lector objetivo no tiene porque tener los mismos conocimientos acerca de la materia que se trata, es decir, existen artículos destinados a todos los públicos con una dificultad de comprensión adaptada para cualquier lector, pero también existen textos cuya comprensión requiere de conocimientos previos acerca de la materia.

El problema viene cuando un paciente, usuario o profesional quiere acceder a un texto, ya sea para su consulta o bien para recomendar su lectura a un paciente en el caso de un médico, por ejemplo, y desconoce cuál es el artículo más adecuado en función de la dificultad de éste y la capacidad de comprensión del lector objetivo. Tal y como se ha comentado antes, es de gran importancia que el paciente sea informado debidamente y para ello sería necesario que los artículos disponibles en la web de Salupedia fueran catalogados en función de su legibilidad lingüística, facilitando la tarea del profesional de la salud que ha de proporcionarle un texto adecuado a su capacidad de comprensión.

Otro problema es que, aunque existan proyectos como Salupedia, es posible que los resultados proporcionados por los buscadores no especializados en medicina, los cuales resultan de uso más extendido, no proporcionen información médica con una legibilidad lingüística adaptada al público general.

[1] http://www.elmundo.es/elmundosalud/2007/07/12/biociencia/1184255594.html
[2] http://www.ub.edu/senesciencia/noticia/empoderamiento-del-paciente/
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1173379/
[4] http://www.salupedia.org/quees.php

Por estas razones, el estudio de la legibilidad de un texto médico puede resultar muy interesante para el nuevo perfil de paciente empoderado y para facilitar la relación que se establece entre un paciente y su médico.

1.2. Distinción entre legibilidad lingüística y tipográfica:

A continuación, se va a clarificar la diferencia conceptual del término legibilidad lingüística, que se utiliza a lo largo del presente trabajo, y legibilidad tipográfica. Aunque en algunas publicaciones se utiliza el término legibilidad para referirse a ambos conceptos, a menudo resulta necesario clarificar si se refiere a legibilidad lingüística o bien legibilidad tipográfica.

- **Legibilidad lingüística:** Este término es equivalente a hablar de la dificultad que presenta el texto para ser comprendido. Depende de factores como la longitud de las palabras o su frecuencia de uso, la longitud de las frases y la estructura de estas, puesto que una frase simple con una estructura sencilla resulta más comprensible. Además, depende también de factores intrínsecos al lector, como pueden ser su formación, su habilidad lectora o su conocimiento del idioma en que está escrito el texto.

  A continuación, se muestra un ejemplo de dos frases que significan lo mismo, pero tienen distinta legibilidad lingüística:

  "*El petirrojo voló y se detuvo sobre la rama del pino"*
  Legibilidad lingüística alta (fácil de comprender)

  "El ejemplar de Erithacus rubecula surcó el cielo en una trayectoria curvilínea para acabar posándose sobre la ramificación de un Pinus halepensis"
  Legibilidad lingüística más baja (Algo más difícil de comprender)

- **Legibilidad tipográfica:** Tiene que ver con la dificultad de lectura de un texto en base a características formales, como por ejemplo la tipografía, el color de fondo, la distribución de los parágrafos, el sangrado, el uso correcto de signos de puntuación, etc. En este campo son relevantes los estudios de Tinker. (Tinker, 1963)

La medición de la legibilidad lingüística de un texto médico es uno de los objetivos del presente trabajo de fin de grado, siempre basando el estudio en valores objetivos, que nos permitan establecer una escala de complejidad adaptada para las muestras de las que se dispone.

2. Objetivos:

Con el objetivo de solucionar los problemas planteados en el *subapartado 1.1*, se va a desarrollar un software que sea capaz de recibir un texto médico y que, tras aplicar una serie de algoritmos, permita la obtención de un nivel de dificultad automáticamente.

El principal objetivo del presente trabajo de fin de grado **es el desarrollo de un software capaz de establecer un nivel de dificultad en la comprensión de un texto de cualquier ámbito del campo de la salud y** a continuación utilizar dicho software para **evaluar los resultados devueltos por un buscador de uso general para un conjunto de palabras clave** (enfermedades de interés general). Para ello se va a desglosar la realización del proyecto en una serie de hitos u objetivos intermedios que permitan finalmente alcanzar dicho propósito.

- Primeramente, se realizará un análisis intensivo del estado del arte (*apartado 3)*, con el fin de determinar cuáles son las técnicas utilizadas hoy en día para asignar un valor de legibilidad lingüística a un texto (*subapartado 3.1*) y se determinará cuáles son las herramientas existentes (*subapartado 3.2*) que mejor se adaptan para el problema planteado.
- Puesto que se dispone de una muestra de textos de ámbito sanitario con un nivel de complejidad asignado por un especialista, se probarán las herramientas existentes sobre esas muestras (*subapartado 5.1*) y se analizará el resultado proporcionado (*subapartado 5.2*) con el objetivo de determinar que herramienta y algoritmo proporciona unos resultados que se ajusten lo máximo posible a la dificultad establecida por el experto en dichos textos.
- A continuación, se desarrollará un software que recibiendo una URL cualquiera realice primeramente un filtrado de la información relevante y a continuación un análisis de la legibilidad lingüística del texto, proporcionando un valor de dificultad en un índice dado (*subapartado 5.3*).
- El siguiente objetivo será determinar cuáles son las siete búsquedas relacionadas con medicina que más realizan los usuarios de Internet (*subapartado 4.3*). Seguidamente, se obtendrá una lista con 10 artículos que un buscador general devuelve a realizar una búsqueda con las palabras clave determinadas anteriormente.
- Finalmente, utilizando el software desarrollado, se obtendrá la complejidad de cada uno de estos artículos y se plasmarán los resultados en una tabla (*subapartado 5.4*) que permita evaluar si los resultados devueltos por buscadores proporcionan resultados con un nivel de complejidad accesible para cualquier usuario medio.

3. Estado del arte

3.1. Medida de la complejidad de un texto

Existen diferentes enfoques desde los que se puede abordar el problema de la determinación de la complejidad de un texto dado de forma automática. Históricamente se han utilizado fórmulas para estimar la complejidad de un texto. Estas fórmulas se basan en el análisis de variables como la longitud de las frases o el número de sílabas que tienen las palabras que las componen.
Con los avances actuales en la informática y el auge de técnicas de aprendizaje automático, una posible aproximación para resolver el problema que se plantea sería la elaboración de un software basado en estas técnicas. Sin embargo, el desarrollo de un programa de este tipo escapa al alcance del presente proyecto ya que no se dispone de la cantidad necesaria de muestras para realizar el entrenamiento, y en la actualidad, al menos para el idioma español, no existen (o no se han encontrado) proyectos basados en estas técnicas.

3.1.1. Fórmulas para calcular la legibilidad lingüística en inglés

Las primeras investigaciones en este campo que fueron ampliamente aceptadas fueron las de Rudolf Flesch. A continuación, se realiza una breve revisión de las fórmulas inglesas más populares (puestos que existen más de 200 (DuBay, 2004)) y su evolución hasta la actualidad:

The Flesch Reading Ease Formula (Flesch, 1948): es una fórmula ampliamente usada en sus inicios y desarrollada por Rudolf Flesch, la cual evalúa la dificultad de un texto escrito en inglés en una escala de 0 a 100 (aunque es posible obtener valores fuera de este rango). El índice que propone es únicamente válido para lectores nativos y se asocia la puntuación obtenida con un nivel de escolarización estadounidense mínimo para la comprensión del texto. En la tabla 1 se muestra el nivel de escolarización asociado a cada rango de valores.

$$206.835 - \left(1.015 \times \frac{Total\ palabras}{Total\ frases}\right) - \left(84.6 \times \frac{Total\ sílabas}{Total\ palabras}\right)$$

(Kincaid, Fishburne, Rogers, & Chissom, 1975)

Puntuación	Nivel de escolarización
De 90 a 100	5º curso
De 80 a 90	6º curso
De 70 a 80	7º curso
De 60 a 70	8º y 9º curso
De 50 a 60	De 10º a 12º curso (Instituto)
De 30 a 50	Universidad
De 0 a 30	Graduado universitario

Tabla 1 – Interpretación de la puntuación de la fórmula de Flesch[5]. Traducción propia al español.

[5] http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml

The Flesch-Kincaid Grade Level (1976): En el año 1976, a petición de la Marina de los EE. UU., Peter Kincaid y su equipo desarrolló una fórmula basada en la "Reading Ease Formula" de Rudolf Flesch. En un principio la fórmula se utilizó para evaluar la dificultad de los manuales técnicos de la marina y al poco tiempo se convirtió en un estándar y se utiliza con frecuencia en sectores diversos, como la enseñanza o para asegurar una legibilidad lingüística en las pólizas de seguros que permita que cualquier persona, con conocimientos básicos, entienda el texto correctamente. Esta fórmula se basa en las mismas variables que la original de Flesch, pero están correladas inversamente, es decir, si se recibe una puntuación alta en la primera, deberá recibirse una puntuación baja en la segunda.

$$\left(0.39 \times \frac{Total\ palabras}{Total\ frases}\right) + \left(11.8 \times \frac{Total\ sílabas}{Total\ palabras}\right) - 15.59$$

(Kincaid et al., 1975)

La fórmula devuelve un número que indica el curso de escolarización (respecto al sistema educativo de los EEUU) necesario para comprender el texto, aunque de forma teórica el valor mínimo que puede dar es de -3.4, el cual se correspondería con el texto más fácil posible (frases de una palabra con palabras monosílabas).

The Gunning's Fog index (Gunning, 1968): Fórmula creada por Robert Gunning en 1952 (El libro de la referencia se trata de una edición posterior). Es especialmente útil para ayudar a escritores y personas cuya profesión no es específicamente la de escribir, pero representa una necesidad, como empresarios, ingenieros, científicos, etc. Esta fórmula permite medir la claridad de un texto y determinar cuan sobrecargado está de complejidad innecesaria para transmitir el mensaje deseado.

$$0.4 \times \left[\left(\frac{Total\ palabras}{Total\ frases}\right) + 100 \times \left(\frac{Palabras\ complejas}{Total\ palabras}\right)\right]$$

Es una fórmula muy sencilla de aplicar (incluso a mano) y con un coste computacional más reducido que las anteriores, ya que trabaja con pocos decimales. La forma de interpretar los resultados es la misma que en la de Flesch-Kincaid, es decir, el número que se obtiene se corresponde con el curso de escolarización necesario para comprender el texto, según el sistema educativo estadounidense [6] [7].

Dale-Chall readability (Dale & Chall, 1948): Esta fórmula estima la legibilidad lingüística de un texto a partir de una lista de 763 palabras que el 80% de estudiantes de cuarto curso es capaz de entender. Resulta útil para niños de más de cuarto curso y adultos. En su segunda versión, actualizada en 1995, la lista de palabras se alarga hasta las 3000:

$$0.1579\left(\frac{palabras\ difíciles}{Total\ palabras} \times 100\right) + 0.0496\left(\frac{Total\ palabras}{Total\ frases}\right)$$

Puntuación	Nivel de escolarización
4.9 e inferior	4º curso e inferior
De 5.0 a 5.9	Cursos 5º - 6º
De 6.0 a 6.9	Cursos 7º - 8º
De 7.0 a 7.9	Cursos 9º - 10º
De 8.0 a 8.9	Cursos 11º - 12º
De 9.0 a 9.9	Cursos 13º - 15º (Universitario)
10 y superior	Cursos 16º y superior (Graduado universitario)

[6] http://www.impact-information.com/impactinfo/newsletter/plwork08.htm
[7] https://readable.io/content/the-gunning-fog-index/

Tabla 2 – Interpretación de la fórmula de Dale-Chall 1995 (Dale & Chall, 1948). Traducción propia al español.

Coleman-Liau Index (Coleman & Liau, 1975): Los autores de la fórmula defienden que la longitud de las palabras en cuanto a número de letras es mejor predictor de la legibilidad lingüística que la longitud de las palabras en cuanto a sílabas, la cual resulta además más costosa de calcular.

$$CL = 0.0588L - 0.296S - 15.8$$

Donde **L** es el número medio de letras por cada 100 palabras y **S** el número medio de frases por cada 100 palabras. El resultado de la fórmula estima el número de años de escolarización necesarios para comprender el texto, respecto al sistema educativo estadounidense.

Estas fórmulas son algunas de las más populares debido a su simplicidad y a que proporcionan una buena correlación con los textos que se utilizaron de muestra. Sin embargo, no proporcionan resultados correctos al aplicarse en un texto en español, debido a las diferencias intrínsecas que existen entre ambos idiomas como la longitud media de las palabras o de las frases, es por este motivo que se utilizan fórmulas propias o adaptadas de las inglesas para la obtención de la legibilidad lingüística de un texto escrito en español.

3.1.2. Fórmulas para calcular la legibilidad lingüística en español

Seguidamente se han revisado las fórmulas para calcular la legibilidad lingüística de un texto en español que pueden resultar de utilidad para este proyecto:

Fórmula de Spaulding (Spaulding, 1956): La primera fórmula para calcular la legibilidad lingüística de un texto en español. Puede ser vista como la adaptación al español de la fórmula de Dale-Chall propia de la lengua inglesa (Gala, Rapp, & Bel-Enguix, 2014), ya que también depende de una lista, en este caso 1500 lemas frecuentes en el castellano y fáciles de comprender.

$$D = 1.609(L) + 331.8(R) + 22.0$$

D es la dificultad del texto, **L** es la longitud media de la frase, **R** representa la densidad de palabras raras, es decir, que no están presentes en la lista. El valor de D se interpreta según la *tabla 3*.

Índice	Dificultad
0 – 40	Primeras enseñanzas y materiales muy simplificados
40 – 60	Muy fácil
61 – 80	Fácil
81 – 100	Moderadamente difícil
101 – 120	Difícil
121 o más	Excepcionalmente difícil

Tabla 3 – Interpretación resultado Fórmula de Spaulding[8].

[8] https://legible.es/blog/spaulding/

Índice Fernández Huerta (Fernández Huerta, 1959): Se trata de una fórmula para medir la legibilidad lingüística creada por José Fernández Huerta en 1959 basándose en la fórmula de Flesch.

$$L = 206.84 - 0.60P - 1.02F$$

En esta fórmula, ***L*** representa la legibilidad lingüística del texto, en una escala en la que valores más altos significan una mayor legibilidad lingüística. La letra ***P*** es el promedio de sílabas por palabra y la ***F*** la media de palabras por frase[9].

L	Nivel	Grado escolar
90 – 100	Muy fácil	4º grado
80 – 90	Fácil	5º grado
70 – 80	Algo fácil	6º grado
60 – 70	Normal (para adulto)	7º u 8º grado
50 – 60	Algo difícil	Preuniversitario
30 – 50	Difícil	Cursos selectivos
0 – 30	Muy difícil	Universitario (especialización)

Tabla 4 – Interpretación resultado fórmula Fernández Huerta[8].

Legibilidad µ: Permite calcular la legibilidad lingüística de un texto, pero utiliza variables distintas al resto de fórmulas, ya que, además de tener en cuenta el total de palabras del texto, también entran en juego otras variables.

$$\boldsymbol{\mu = \left(\frac{n}{n-1}\right)\left(\frac{\overline{x}}{\sigma^2}\right) \times 100}$$

µ es el índice de legibilidad, **n** es el total de palabras en el texto, $\overline{x}$ es la media del número de letras por palabra y σ^2 la varianza del número de letras por palabra.

Índice	Facilidad de lectura
91 - 100	Muy fácil
81 - 90	Fácil
71 - 80	Un poco fácil
61 - 70	Adecuado
51 - 60	Un poco difícil
31 - 50	Difícil
0 - 30	Muy difícil

Tabla 5 – Interpretación del valor µ.

Índice Flesch-Szigriszt ("Fórmula de perspicuidad")(1993): Se trata de una readaptación de la fórmula de Flesch al español realizada por Francisco Szigriszt Pazos en su tesis doctoral (Szigriszt Pazos, 2001). Se conoce también como índice Szigriszt-Pazos o fórmula de perspicuidad. Según la RAE, en su tercera acepción:

> **Perspicuidad**: adj. Dicho del estilo: inteligible (‖ que puede ser entendido).

[9] https://legible.es/blog/lecturabilidad-fernandez-huerta/

Es decir, en este caso se utiliza la palabra perspicuidad a modo de sinónimo del término legibilidad lingüística. Es la **fórmula de referencia para el cálculo de la legibilidad lingüística en español**.

$$P = 206.835 - \frac{62.3S}{W} - \frac{W}{F}$$

P es la perspicuidad, **S** el total de sílabas, **W** la cantidad de palabras y **F** el número de frases[10].

Puntos	Estilo	Tipo de publicación	Estudios
0 a 15	Muy difícil	Científica, filosófica	Titulados universitarios
16 a 35	Árido	Pedagógica, técnica	Selectividad y estudios universitarios
36 a 50	Bastante difícil	Literatura y divulgación	Cursos secundarios
51 a 65	Normal	Los media	Popular
66 a 75	Bastante fácil	Novela y revista femenina	12 años
76 a 85	Fácil	Para kioscos	11 años
86 a 100	Muy fácil	Cómics, tebeos y viñetas	6 a 10 años

Tabla 6 – Interpretación Índice Flesch-Szigriszt. Se han omitido algunas columnas de la tabla original innecesarias para la interpretación de los resultados. (Szigriszt Pazos, 2001)

Índice INFLESZ (Barrio Cantalejo, 2007): Se trata de la misma fórmula que la de Flesch-Szigriszt pero con la creación de una nueva escala, la escala INFLESZ, ya que la escala del nivel de Perspicuidad propuesta por Szigriszt para interpretar su fórmula precisa adaptación, al haber sido realizada con una muestra insuficiente, no representativa ni aleatoria de textos (Barrio Cantalejo, 2007). La escala que propone la autora es la siguiente:

Puntos	INFLESZ
0 – 40	Muy difícil
45 – 55	Algo difícil
60 – 65	Normal
70 – 80	Bastante fácil
85 – 100	Muy fácil

Tabla 7 – Escala INFLESZ para la fórmula de Flesch-Szigriszt.(Barrio Cantalejo, 2007)

La granularidad queda reducida de siete a 5 niveles de dificultad diferentes respecto a la escala propuesta por Szigriszt. La autora obtuvo esta nueva escala tomando una muestra representativa de textos escritos en español, compuesta de 210 textos de tipos diferentes.

En la *figura 1* se muestra una imagen comparativa de la escala INFLESZ, la escala de Szigriszt y la de Flesch:

[10] https://legible.es/blog/perspicuidad-szigriszt-pazos/

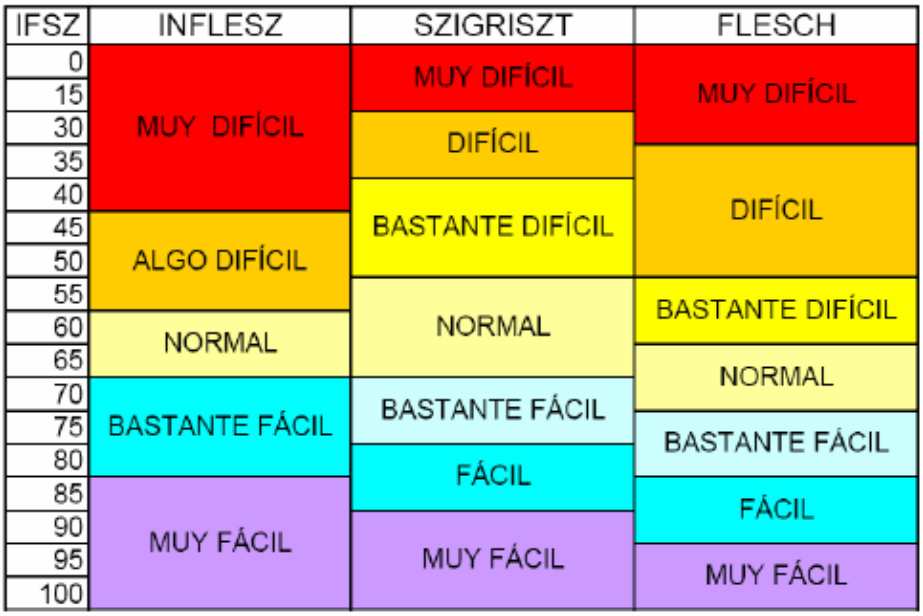

Figura 1 - Comparación de la escala INFLESZ, la escala de Szigriszt y la escala Flesch. (Barrio Cantalejo, 2007)

Actualmente la solución más extendida consiste en la utilización de la fórmula de Flesch-Szigriszt y la escala INFLESZ, la cual proporciona los mejores resultados dentro de los límites existentes en esta metodología.

3.2. Algoritmos de legibilidad lingüística:

En este apartado se realiza una investigación para determinar cuáles son las herramientas existentes, de uso local u online, para calcular la legibilidad lingüística de los artículos proporcionados. Además, se describe cada una de ellas y se elabora una tabla comparativa.

3.2.1. INFLESZ

INFLESZ es un sencillo programa desarrollado como resultado de la tesis doctoral de Inés Mª Barrio Cantalejo, el cual permite evaluar la legibilidad de un texto escrito en español. Este software es capaz de analizar tanto archivos como fragmentos de texto, además se trata de software freeware, por lo que se puede descargar de forma gratuita. INFLESZ está programado en C++ y compilado para entornos Windows.

Este programa Calcula 9 parámetros útiles para evaluar la legibilidad de un texto escrito en español[11]. Son los siguientes:

- Palabras.
- Sílabas.
- Frases.

[11] https://legibilidad.blogspot.com/2015/01/el-programa-inflesz.html

- Promedio sílabas / palabra.
- Promedio palabras / frase.
- Índice Flesch-Szigriszt. (“Fórmula de perspicuidad”)
- Grado en la Escala Inflesz.
- Correlación Word.
- Índice Flesch-Fernández Huerta.

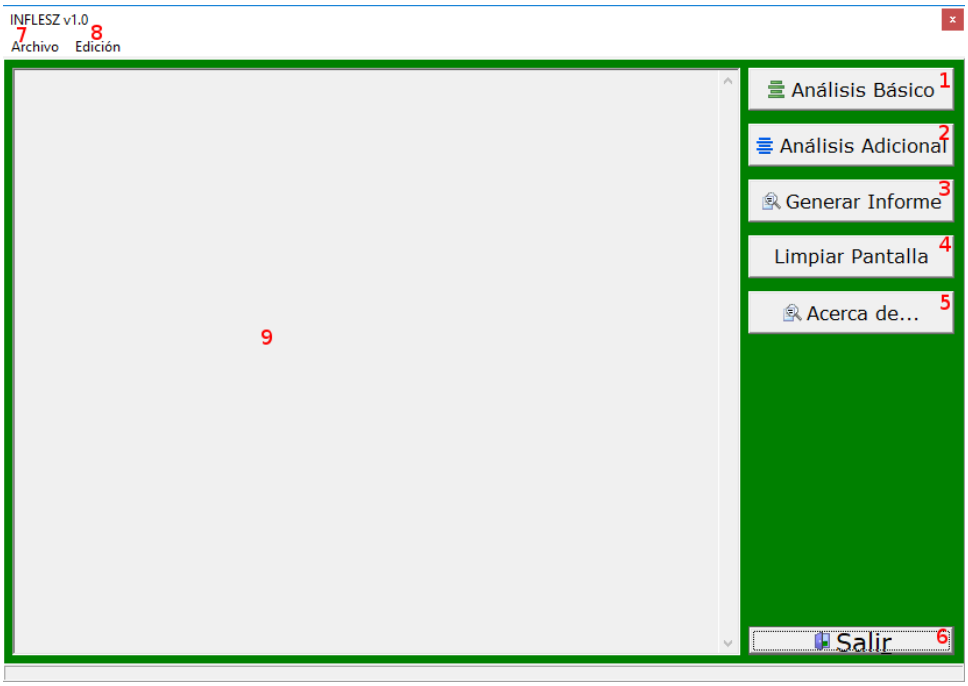

Figura 2 – Interfaz del programa INFLESZ.

[1] Análisis Básico: Ofrece información básica del texto analizado, como número de palabras entre otros. Además, calcula el índice Flesch-Szigriszt y el grado de dificultad en la escala INFLESZ.
[2] Análisis Adicional: Calcula la correlación que devolvería el editor de textos Word para el texto analizado, y también el índice Fernández Huerta.
[3] Generar Informe: permite generar un fichero .HTML con el texto analizado y la información del Análisis Básico [1] organizada en una tabla.
[4] Limpiar Pantalla: Elimina el texto del campo [9]
[5] Acerca de: Muestra la ayuda del programa.
[6] Salir: Cierra el programa.
[7] Archivo: Permite abrir (ficheros .TXT) o guardar (ficheros .IFZ)
[8] Edición: Típicos comando de edición.
[9] Campo de texto, donde se introduce el fragmento que se desea analizar.

El programa utiliza la escala INFLESZ, la cual establece 5 niveles de dificultad:

PUNTOS	GRADO	TIPO DE PUBLICACIÓN
< 40	MUY DIFÍCIL	UNIVERSITARIO, CIENTÍFICO
40-55	ALGO DIFÍCIL	BACHILLERATO, DIVULGACIÓN CIENTÍFICA, PRENSA ESPECIALIZADA

55-65	NORMAL	E.S.O., PRENSA GENERAL, PRENSA DEPORTIVA
65-80	BASTANTE FÁCIL	EDUCACIÓN PRIMARIA, PRENSA DEL CORAZÓN, NOVELAS DE ÉXITO
> 80	MUY FÁCIL	EDUCACIÓN PRIMARIA, TEBEOS, CÓMIC

Tabla 8 – Escala INFLESZ. Documentación del software INFLESZ

3.2.2. Professional Spanish Lexile Analyzer

Esta herramienta ofrecida por MetaMetrics mide la complejidad de un texto, separándolo en partes y estudiando sus características, como la longitud de las frases y la frecuencia de las palabras, las cuales representan el desafío sintáctico y semántico que el texto representa para el lector. El resultado es la complejidad del texto, expresado en la escala Lexile, junto con información del número de palabras, la longitud media de las frases y la frecuencia logarítmica media. Es capaz tanto de analizar textos en inglés como en español.

Spanish Lexile measure – Este valor indica la dificultad de lectura del texto en términos de dificultad semántica y complejidad sintáctica. El rango de valores normal obtenidos en la escala Lexile de El Sistema va desde 200L a 1700L (de más fácil a más difícil), aunque las medidas de Lexile españolas reales pueden ir en un rango de menos de 0L hasta valores superiores a 2000L [12].

3.2.3. Legible.es

Se trata de una aplicación web desarrollada por Alejandro Muñoz Fernández, la cual permite calcular una serie de fórmulas de legibilidad lingüística validadas para el español. Las fórmulas que calcula son:

- Índice Fernández Huerta
- Fórmula de comprensibilidad de Gutiérrez de Polini
- Fórmula de Crawford
- Índice de perspicuidad de Szigriszt-Pazos
- Escala INFLESZ
- Legibilidad µ

Además de calcular estas fórmulas ofrece una estimación del tiempo de lectura del texto procesado, estadísticas del texto (como el número de palabras, frases o párrafos, entre otros), una tabla con la frecuencia de aparición de las letras y otra de las palabras en el texto.

La interfaz de la aplicación web, visible en la *Figura 3,* es muy sencilla, únicamente dispone de un formulario formado por un campo de texto y un botón. En el campo de texto se tiene que introducir un fragmento de texto o bien la URL de la web que se desee analizar, y a continuación pulsar sobre el botón “Analizar”.

[12] https://lexile.com/tools/spanish-lexile-analyzer/step-5-analyze-text-and-get-results/

Figura 3 – Interfaz de la aplicación web de legible.es.

El autor de legible.es ofrece de forma libre (su licencia es GPLv3) el código fuente de los scripts de Python de la aplicación, para que el usuario pueda descargarlos y utilizarlos o modificarlos de forma local.

3.2.4. Legibilidadmu.cl

Se trata de la aplicación web oficial de los autores de la fórmula de legibilidad µ, Miguel Muñoz Baquedano y José Muñoz Urra. Está programada utilizando el complemento de Adobe Flash Player y se necesita para poder ejecutarla disponer de dicho complemento en el navegador.
La herramienta, además de proporcionar el valor de la legibilidad µ y el valor de dificultad asociado en la escala, proporciona también una serie de datos de interés para el usuario, como el total de palabras, el número de caracteres, la media de caracteres, la desviación típica o la varianza.

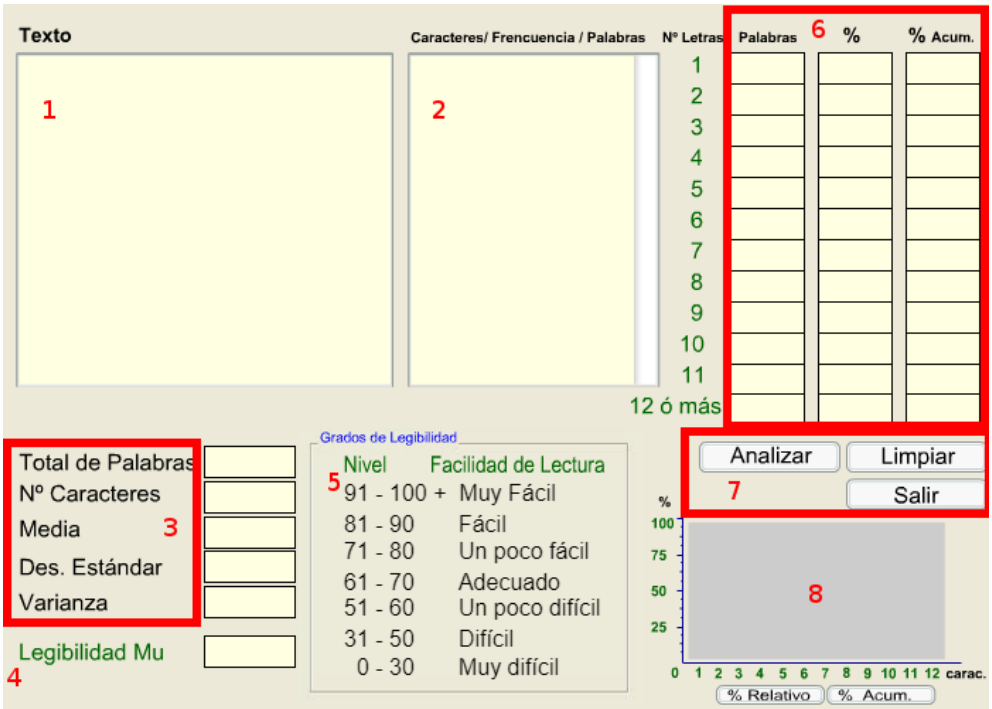

Figura 4 – Interfaz de la aplicación Flash de legibilidadmu.cl

[1] Es el campo donde se introduce el texto que se va a analizar.
[2] Contiene información de cada palabra que aparece en el texto: el número de caracteres que tiene y la frecuencia con que se repite
[3] Conjunto de parámetros que calcula el analizador.
[4] Valor del índice µ
[5] Escala del índice de legibilidad µ
[6] Recuento de palabras en función del número de letras que poseen. Porcentaje de aparición y porcentaje acumulado.
[7] Botones. "Analizar" calcula la legibilidad µ para el fragmento de texto introducido y rellena todos los campos. "Limpiar" elimina el fragmento de texto y vacía todos los campos. "Salir" cierra la aplicación.
[8] Gráfica que representa los valores de la segunda y tercera del apartado [6].

3.2.5. Tabla comparativa

En la *tabla 9* se comparan algunas características de las diferentes herramientas que se han presentado:

	Herramienta online	**Análisis de texto por URL**	**¿Qué fórmulas soporta?**
INFLESZ v1.0	No	No	- Índice Flesch-Fernández Huerta - Índice Flesch-Szigriszt - Escala INFLESZ
Legibilidadmu.cl	Sí	No	- Legibilidad µ
Legible.es	Sí	Sí	- Índice Fernández Huerta - Fórmula de Gutierrez - Fórmula de Crawford - Índice Szigriszt-Pazos (Flesch-Szigriszt) - Escala INFLESZ - Legibilidad µ
Professional Spanish Lexile Analyzer	Sí	No	- Fórmula propia

Tabla 9 – Tabla comparativa de las diferentes herramientas analizadas. Elaboración propia.

3.3. ¿Cómo buscan información sanitaria los pacientes?

La información sanitaria es por lo general suministrada por los especialistas de la salud, es decir, el personal médico, sanitario y el farmacéutico. No obstante, hoy en día, debido a la generalización del uso de Internet y por la comodidad que supone el acceso inmediato a la información, es común el uso de buscadores web (principalmente Google) para obtener información sanitaria, ya sea previa o posterior al diagnóstico médico. Según un estudio del Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información (ONTSI)[13], dependiente del Ministerio de Industria y gestionado por Red.es, el 60.5% de la población española utiliza Internet para consultar información sobre salud.

[13] http://www.ontsi.red.es/ontsi/es/estudios-informes/los-ciudadanos-ante-la-e-sanidad-opiniones-y-expectativas-de-los-ciudadanos-sobre-

Normalmente, los pacientes que consultan información en la web lo hacen a través de buscadores genéricos, como podría ser Google, Bing o Yahoo. El paciente que necesita asesoramiento o información médica acerca de alguna enfermedad, tratamiento o síntomas accede a algún buscador e introduce su consulta directamente, en lugar de consultar buscadores especializados en información de la salud. Actualmente existen diversos buscadores de este tipo, como por ejemplo:

- **Salupedia:** Es un proyecto colaborativo con el objetivo de crear un lugar donde profesionales (médicos, enfermeros, psicólogos, etc.) recomienden contenidos, ya existentes en la red, a pacientes, familiares y ciudadanos en general, asegurando en la medida de lo posible que se trata de información fiable[14].
- **MedlinePlus**: Es el sitio web de los Institutos Nacionales de la Salud para pacientes, familiares y amigos. Producida por la Biblioteca Nacional de Medicina de los Estados Unidos. Brinda información sobre enfermedades, afecciones y bienestar en un lenguaje fácil de leer. Ofrece información confiable y actualizada en todo momento, en cualquier lugar y de forma gratuita [15].
- **Forum Clinic:** Es un programa interactivo para pacientes destinado a que aumenten su grado de autonomía con respecto a su salud, utilizando las oportunidades que brindan las nuevas tecnologías. Aporta información rigurosa, útil, transparente y objetiva sobre la salud, al tiempo que favorece la participación activa de los pacientes y sus asociaciones. Pretende contribuir a que los ciudadanos mejoren el conocimiento sobre la salud, la enfermedad y sus condicionantes, y sobre la eficacia y seguridad de los tratamientos curativos y preventivos disponibles, para que puedan implicarse en las decisiones clínicas que les afectan[16].
- **HONcode Search:** Se trata de una web para buscar información médica confiable. Fue fundada para "fomentar la difusión de información sanitaria de calidad para pacientes, profesionales y para el público en general y facilitar el acceso a los últimos datos médicos más relevantes a través del uso de Internet"[17].

[14] http://www.salupedia.org/quees.php
[15] https://medlineplus.gov/spanish/aboutmedlineplus.html
[16] http://www.forumclinic.org/es/general/%C2%BFqu%C3%A9-es-forumcl%C3%ADnic
[17] http://www.hon.ch/HONcode/Search/search_sp.html

4. Materiales y métodos

En este apartado se explica cómo se ha llevado a cabo el desarrollo del proyecto, indicando las herramientas que se han utilizado, los recursos de los que se disponía y el procedimiento seguido para alcanzar los resultados.

4.1. Herramientas de desarrollo

A continuación, se va a hacer una descripción de las herramientas que se han utilizado en el desarrollo del programa. Desde el entorno de desarrollo, pasando por el lenguaje de programación que se ha utilizado y finalmente una breve explicación de las librerías de las que se ha hecho uso.

4.1.1. Notepad++

Se trata de un editor de texto gratuito, con licencia GNU, que soporta una gran variedad de lenguajes de programación diferentes, permitiendo incluso definir un lenguaje propio al usuario. Se ha elegido este editor de texto debido a su simplicidad y a las siguientes características que lo hacen especialmente útil:

- Resaltado y plegado (ocultar o mostrar bloques de código) de sintaxis.
- Interfaz minimalista y personalizable.
- Soporte para Python (y otros 71 lenguajes más).
- Autoguardado.
- Edición multilínea.
- Soporta uso de macros.
- Posibilidad de utilizar expresiones regulares para encontrar y reemplazar texto.
- Posibilidad de añadir nuevas funcionalidades con un sistema de Plugins.

Figura 5 – Logo de Notepad++

4.1.2. Python

Python es un lenguaje de programación interpretado, de alto nivel, multiparadigma, multiplataforma y de tipado dinámico creado a finales de los ochenta por Guido van Rossum.

Figura 6 – Logo de Python

Cuando se dice que Python se trata de un lenguaje interpretado, significa que no se necesita un proceso de compilación del código fuente para la ejecución del programa. La ejecución del código por el intérprete de Python es más lenta que la ejecución de código ya compilado, ya que el intérprete tiene que analizar cada sentencia del programa cada vez que se ejecuta (particularmente lento en bucles). Sin embargo, el uso de lenguajes de programación interpretados tiene una serie de ventajas como:

- **Ejecución multiplataforma**: ya que el código puede correr en cualquier sistema operativo que disponga del intérprete de Python.
- **Uso de tipos dinámicos**: una misma variable definida en Python puede tomar valores de distinto tipo en tiempo de ejecución.
- Permite **ejecutar código "on the fly"**, sin tener que compilar (Se ejecuta antes, pero no más rápido).
- Por lo general, los **interpretes** de código son **más fáciles de implementar** que los compiladores.

Python es multiparadigma, ya que soporta:

- Orientación a objetos
- Programación imperativa
- Programación funcional
- Reflexión

Un paradigma es una filosofía o enfoque de programación que define una serie de metodologías y abstracciones para facilitar la solución de problemas de la ingeniería de software.

Python es un lenguaje **libre, de código abierto**, que utiliza una licencia propia denominada Python Software Foundation License, compatible con GPL. Otra ventaja de Python es su filosofía, la cual hace **hincapié en utilizar una sintaxis limpia y un código legible**, eliminando el uso de delimitadores de bloques como las llaves ({}) que se utilizan en otros lenguajes y las sustituye por espacios o tabuladores, el uso de los cuales obliga a los desarrolladores a elaborar código legible y fácil de interpretar.

Python incluye un modo interactivo que permite testear pequeñas porciones de código directamente en el intérprete, esto unido a la gran cantidad de librerías estándar y su elegante sintaxis hacen de Python un lenguaje ideal para el desarrollo de prototipos.

4.1.3. urllib Package

urllib es un paquete de la librería estándar de Python, el cual integra varios módulos que permiten trabajar con URL's[18]. Estos módulos son:

[18] https://docs.python.org/3/library/urllib.request.html

- urllib.request → se utiliza para abrir y leer URL's.
- urllib.error → contiene las excepciones generadas por urllib.request.
- urllib.parse → permite parsear (analizar sintácticamente) URLs.
- Urllib.robotparser → permite parsear ficheros robots.txt

De estos módulos, el único que se va a utilizar para el desarrollo del software propuesto es urllib.request, el cual define una serie de funciones y clases que permiten abrir URLs (principalmente HTTP) y soporta gestión de autentificación, redireccionamiento, cookies y más. Dentro de este módulo resulta de especial interés el método urlopen el cual se va a documentar de forma resumida a continuación:

```
urlopen(url, data=None, [timeout, ]*, cafile=None, capath=None, cadefault=False
, context=None)
```

> **url:** String con la URL (u objeto *Request*)
> **data**: Un objeto que contiene información adicional para el servidor. None para no enviar nada (valor por defecto).
> **timeout:** es opcional, especifica un valor de timeout para operaciones como el intento de conexión.
> **cafile y capath**: se utilizan para indicar una o un conjunto autoridades de certificación (CA) para utilizar HTTPS
> **cadefault**: actualmente es ignorado.
> **context**: si se especifica, es un objeto SSLContext que configura las opciones de SSL.

Recupera la información de una URL, mediante un request si se trata de un servidor HTTP. Es la función principal que se va a utilizar.
Cuando la URL corresponde a un recurso HTTP, la función devuelve un objeto http.client.HTTPResponse modificado. Se puede recuperar la información contenida entre las etiquetas <body></body> de la respuesta HTTP utilizando el método read() de la clase HTTPResponse.

4.1.4. readability-lxml

readability-lxml es una librería gratuita con licencia Apache 2.0 que, dado un documento HTML, extrae el texto principal y limpia la información sobrante, como barras de navegación o el pie de página. Es capaz de detectar también el título del artículo.[19]

4.1.5. BeautifulSoup

Es una librería de Python gratuita publicada bajo licencia MIT, que se utiliza para extraer información de ficheros HTML y XML (incluso los que tienen un marcado incorrecto), resulta útil para el Web Scraping, es decir, la utilización de software para la extracción de información de sitios web. Hay tres características principales que lo hacen especialmente potente:

1. Proporciona una serie de métodos que permiten **navegar, buscar y modificar un árbol de parseo** (árbol de jerarquía generado a partir de las etiquetas HTML o XML), permitiendo crear una aplicación para extraer información de una web con poco código y tiempo.
2. BeautifulSoup hace una **conversión automática de los ficheros de entrada a Unicode y convierte los de salida a UTF-8**. De este modo el desarrollador no ha de preocuparse por la codificación, a menos que el documente no especifique la codificación original y BeautifulSoup no sea capaz de detectarla, en cuyo caso es el desarrollador el que debe indicar la codificación del archivo.

[19] https://pypi.python.org/pypi/readability-lxml

3. BeautifulSoup **soporta los parsers** (analizadores sintácticos) **más utilizados, como "lxml" y "html5lib"**, **además de** soportar el parser **incluido en la biblioteca estándar de Python**. La utilización de distintos parsers permite utilizar diferentes técnicas dando prioridad por ejemplo a la velocidad de ejecución sobre la flexibilidad o al revés, en función del parser que se utilice.

4.1.6. Librería legible.es

Es una librería de Python gratuita, publicada bajo licencia GPLv3, que se puede descargar desde el GitHub[20] de su autor. Se utiliza para calcular la legibilidad lingüística de un texto o fragmento dado. Incluye funciones que permiten determinar variables como el número de palabras, de frases, de sílabas o de párrafos. Estas variables son necesarias para el cálculo de los diferentes índices que esta librería es capaz de obtener, estos son:

- Índice Fernández-Huerta
- Índice Gutiérrez
- Índice Szigriszt-Pazos
- Índice INFLESZ
- Legibilidad µ
- Fórmula de Crawford

Además, también dispone de funciones que permiten interpretar los diferentes índices utilizando las escalas que los autores han definido para cada una de ellas.

4.2. Muestras de texto y método de análisis

Se dispone de una selección de textos de ejemplo que han sido catalogados en cuanto a su nivel de dificultad de comprensión por un experto en el campo de la salud. Los textos aparecen con el título del artículo y clasificados según su dificultad en la *tabla 10.*

Artículo	Dificultad asignada por profesionales de la salud
"La artritis reumatoide"	Difícil
"Vía biliar: Cólico biliar, colecistitis, colangitis y coledocolitiasis ¿Cómo distinguirlas?"	Difícil
"Guía del paciente a los marcadores tumorales"	Difícil
"Artritis idiopática juvenil"	Difícil
"Chikungunya"	Difícil
"Enfermedad por el virus de Zika"	Difícil
"Artritis reumatoide"	Media
"Colecistitis aguda"	Media
"Decálogo de actuación en los colegios ante las alergias"	Media
"¿Cómo se evalúa el dolor en los niños?"	Media
"La hipoglucemia"	Media
"¿Qué es la artritis reumatoide?"	Fácil

[20] https://github.com/amunozf/legibilidad

“Colecistitis”	Fácil
“Dieta para pacientes con Enfermedad inflamatoria intestinal”	Fácil
“Tu hijo quiere una mascota: analiza los pros y los contras”	Fácil
“Consejos para el niño que lleva tupper a la escuela”	Fácil

Tabla 10 – Listado de textos de muestra. Los textos están agrupados por dificultad, PERO NO ORDENADOS por dificultad dentro de cada grupo. Elaboración propia.

Como uno de los objetivos de la investigación es la clasificación de los textos para facilitar el acceso al contenido de la web de Salupedia a los usuarios en función de sus conocimientos, se ha definido una escala de tres niveles que resulta suficiente para clasificar los textos de muestra de los que se dispone. A lo largo de la memoria se va a utilizar el siguiente código de colores para indicar el nivel de legibilidad lingüística de cada texto:

- Verde → Fácil
- Azul → Media
- Rojo → Difícil

Para analizar estos textos se han utilizado las herramientas descritas en el *apartado 3.2*. Primero se ha calculado la legibilidad lingüística para cada una de las muestras con cada una de estas herramientas (*apartado 5.1*) y a continuación, se ha realizado un estudio de los resultados haciendo uso de Matlab.
Para evaluar la bondad de los resultados proporcionados se ha definido una escala de 3 niveles para cada índice. Los límites de cada nivel se han determinado teniendo en cuenta el valor de la media y la desviación típica de cada grupo de texto (difíciles, medio y fáciles). Conociendo la media y la desviación típica de cada grupo se puede encontrar el punto de corte para estimar los límites de la escala más acertados estadísticamente, para una muestra dada. **Sin embargo, es importante destacar que las escalas obtenidas no están validadas, puesto que la muestra de textos es demasiado reducida y no resulta representativa de la población**. Después de elaborar la escala para cada índice se ha escogido la fórmula que proporciona los resultados más acertados para la muestra de textos que se tiene. Finalmente, ésta ha sido la fórmula utilizada para el desarrollo del software, juntamente con la escala de 3 niveles definida para ésta.

4.3. Selección de las enfermedades que se van a buscar

Se ha realizado una investigación de las enfermedades que los usuarios de Internet españoles buscan con mayor frecuencia. Se ha decidido utilizar como motor de búsqueda Google, puesto que es el más popular actualmente[21]. A partir de la información proporcionada por Google Trends, siete de las enfermedades que más han buscado los usuarios de Google en el año 2016 en España son[22]:

- Cáncer
- Lupus
- Gripe
- Diabetes
- Herpes
- Alzheimer
- Sida

[21] http://www.ebizmba.com/articles/search-engines
[22] http://www.ippok.com/blog/enfermedades-mas-buscadas-google/

El siguiente paso ha sido consultar en Google información sobre estas enfermedades, recogiendo los primeros diez artículos para a continuación evaluarlos con la herramienta que se ha desarrollado. A partir de los resultados obtenidos, se ha elaborado una tabla que permite evaluar si los artículos sugeridos por Google al realizar una búsqueda están adaptados para un público sin demasiados conocimientos acerca de salud y medicina en general.

5. Resultados:

5.1. Testeo de las herramientas:

Para la comparación de las herramientas se dispone del conjunto de textos de la *tabla 10*. Se va a comparar el resultado que proporcionan cada una de las herramientas expuestas en el apartado 4 en cada uno de estos textos de ejemplo. Estableceremos la bondad de los resultados arrojados por estos programas en base a la similitud del resultado proporcionado por el software y el proporcionado por especialistas en el campo de la medicina. En general, se puede asumir que una muestra de texto más grande ofrecerá resultados estadísticamente mejores sobre la legibilidad lingüística de dicho texto, por lo que se va a analizar cada artículo al completo, pero cumpliendo unas condiciones formales que se relatarán más adelante, para adaptar el texto a las herramientas que se van a utilizar. Para obtener resultados coherentes se va a utilizar exactamente el mismo texto de entrada para cada herramienta.

5.1.1. INFLESZ:

A continuación, se calcula la legibilidad lingüística de cada texto mediante la aplicación INFLESZv1.0, por el método Férnandez-Huerta y Flesch-Szigriszt. Los resultados del índice Flesch-Szigriszt se interpretan utilizando la escala INFLESZ (interpretación en la columna "Grado en la escala INFLESZ"). La interpretación del resultado del índice Fernández-Huerta puede obtenerse consultando la *tabla 4*.

Artículo	Índice Férnandez-Huerta	Grado en la escala Férnandez-Huerta*	Índice Flesch-Szigriszt	Grado en la escala INFLESZ
"La artritis reumatoide"	59.73	Algo difícil	54.89	Algo difícil
"Vía biliar: Cólico biliar, colecistitis, colangitis y coledocolitiasis ¿Cómo distinguirlas?"	53.58	Algo difícil	48.33	Algo difícil
"Guía del paciente a los marcadores tumorales"	58.22	Algo difícil	52.91	Algo difícil
"Artritis idiopática juvenil"	43.91	Difícil	38.98	Muy difícil
"Enfermedad por el virus de Zika"	57.42	Algo difícil	52.79	Algo difícil
"Chikungunya"	61.57	Normal (para adulto)	56.81	Normal
"Artritis reumatoide"	52.48	Algo difícil	47.74	Algo difícil
"Colecistitis aguda"	62.93	Normal (para adulto)	58.12	Normal
"Decálogo de actuación en los colegios ante las alergias"	61.30	Normal (para adulto)	56.33	Normal
"La hipoglucemia"	58.39	Algo difícil	53.63	Algo difícil

“¿Cómo se evalúa el dolor en los niños?”	70.89	Algo fácil	66.43	Bastante fácil
“¿Qué es la artritis reumatoide?”	63.01	Normal (para adulto)	58.27	Normal
“Colecistitis”	65.97	Normal (para adulto)	61.18	Normal
“Dieta para pacientes con Enfermedad inflamatoria intestinal”	67.57	Normal (para adulto)	63.11	Normal
“Tu hijo quiere una mascota: analiza los pros y los contras”	65.23	Normal (para adulto)	60.81	Normal
“Consejos para el niño que lleva tupper a la escuela”	71.46	Algo fácil	67.08	Bastante fácil

Tabla 11 – Resultados INFLESZ. Elaboración propia.

*El programa INFLESZ no proporciona la interpretación del valor del índice Fernández-Huerta, no obstante, para la realización de la *tabla 11* se ha completado añadiendo manualmente esta información a partir de la escala Fernández-Huerta, disponible en la *tabla 4*.

5.1.2. Professional Spanish Lexile Analyzer

Para poder probar la herramienta que ofrece MetaMetrics, se han de preparar los textos que va a tomar como entrada según unas normas que se explican en su página web. La herramienta soporta los tipos de textos que se indican a continuación:

Textos que **si se deben** medir:

- Periódicos y **artículos** de revistas
- Libros
- Historias cortas y lecturas
- Pasajes, entrevistas y obras para pruebas
- **Páginas web**

Textos que **no se deben** medir:

- Trabajos escritos del estudiante
- Poesía
- Canciones
- Preguntas de respuesta múltiple
- Texto no escrito en prosa
- Texto sin una puntuación o formato convencional

En la web de Lexile se explica cómo preparar el texto para el análisis, indicando que partes del texto debemos dejar para analizar y cuales se deben eliminar (por ejemplo: URL’s, tablas y gráficos, abreviaciones, frases incompletas, etc.). Los resultados de la *tabla 12* se han obtenido utilizando la herramienta online “Professional Spanish Lexile Analyzer” y los textos utilizados han sido adaptados según las instrucciones de la web por mí, y no por personal cualificado de MetaMetrics, por lo que los resultados obtenidos podrían no ser lo más precisos posible.

Artículo	Puntuación Professional Spanish Lexile Analyzer
“La artritis reumatoide”	1290L
“Vía biliar: Cólico biliar, colecistitis, colangitis y coledocolitiasis ¿Cómo distinguirlas?”	1370L
“Guía del paciente a los marcadores tumorales”	1860L

“Artritis idiopática juvenil”	1510L
“Enfermedad por el virus de Zika”	1260L
“Chikungunya”	1100L
“Artritis reumatoide”	1340L
“Colecistitis aguda”	860L
“Decálogo de actuación en los colegios ante las alergias”	750L
“La hipoglucemia”	1300L
“¿Cómo se evalúa el dolor en los niños?”	1060L
“¿Qué es la artritis reumatoide?”	890L
“Colecistitis”	790L
“Dieta para pacientes con Enfermedad inflamatoria intestinal”	880L
“Tu hijo quiere una mascota: analiza los pros y los contras”	1030L
“Consejos para el niño que lleva tupper a la escuela”	970L

Tabla 12 – Resultados de la herramienta Professional Spanish Lexile Analyzer. Elaboración propia.

5.1.3. Legible.es

El test de *legible.es* proporciona 5 índices para determinar la legibilidad lingüística, además calcula también la fórmula de Crawford (para alumnos de primaria) que permite estimar el número de años de escolarización necesario para comprender el texto. Los resultados de la *tabla 13 y 14* se han obtenido al pasar el test de *legible.es* a cada uno de los textos proporcionados. Se ha dividido la tabla en dos partes para facilitar la visualización de los resultados.

Artículo	**Índice Fernández-Huerta**		**Índice Gutiérrez**		**Índice Szigriszt-Pazos**	
	Valor	Dificultad	Valor	Dificultad	Valor	Dificultad
“La artritis reumatoide”	59.06	Bastante difícil	38.51	Normal	53.92	Normal
“Vía biliar: Cólico biliar, colecistitis, colangitis y coledocolitiasis ¿Cómo distinguirlas?”	53.26	Bastante difícil	36.83	Normal	47.86	Bastante difícil
“Guía del paciente a los marcadores tumorales”	58.14	Bastante difícil	38.43	Normal	52.94	Normal
“Artritis idiopática juvenil”	44.77	Difícil	33.23	Difícil	40.13	Bastante difícil

“Enfermedad por el virus de Zika”	58.55	Bastante difícil	38.18	Normal	53.88	Normal
“Chikungunya”	62.07	Normal	39.77	Normal	57.48	Normal
“Artritis reumatoide”	55.0	Bastante difícil	38.21	Normal	50.38	Normal
“Colecistitis aguda”	64.03	Normal	41.28	Normal	59.2	Normal
“Decálogo de actuación en los colegios ante las alergias”	61.72	Normal	39.63	Normal	56.74	Normal
“La hipoglucemia”	56.9	Bastante difícil	38.21	Normal	51.97	Normal
“¿Cómo se evalúa el dolor en los niños?”	70.97	Bastante fácil	43.82	Normal	66.55	Bastante fácil
“¿Qué es la artritis reumatoide?”	62.04	Normal	40.36	Normal	57.31	Normal
“Colecistitis”	66.81	Normal	43.18	Normal	62.01	Normal
“Dieta para pacientes con Enfermedad inflamatoria intestinal”	70.76	Bastante fácil	42.68	Normal	66.16	Bastante fácil
“Tu hijo quiere una mascota: analiza los pros y los contras”	65.61	Normal	41.51	Normal	60.91	Normal
“Consejos para el niño que lleva tupper a la escuela”	70.08	Bastante fácil	43.7	Normal	65.85	Bastante fácil

Tabla 13 – Resultados Legible.es 1ª parte. Elaboración propia.

Artículo	**Índice INFLESZ***		**Índice legibilidad µ**		**Fórmula de Crawford (años)**
	Valor	Dificultad	Valor	Dificultad	
“La artritis reumatoide”	53.92	Algo difícil	44.59	Difícil	6.0
“Vía biliar: Cólico biliar, colecistitis, colangitis y coledocolitiasis ¿Cómo distinguirlas?”	47.86	Algo difícil	40.15	Difícil	6.3

“Guía del paciente a los marcadores tumorales”	52.94	Algo difícil	47.16	Difícil	5.4
“Artritis idiopática juvenil”	40.13	Algo difícil	46.31	Difícil	7.0
“Enfermedad por el virus de Zika”	53.88	Algo difícil	47.92	Difícil	6.1
“Chikungunya”	57.48	Normal	48.4	Difícil	5.8
“Artritis reumatoide”	50.38	Algo difícil	45.97	Difícil	6.4
“Colecistitis aguda”	59.2	Normal	58.47	Un poco difícil	5.5
“Decálogo de actuación en los colegios ante las alergias”	56.74	Normal	52.16	Un poco difícil	5.6
“La hipoglucemia”	51.97	Algo difícil	50.92	Difícil	6.2
“¿Cómo se evalúa el dolor en los niños?”	66.55	Bastante fácil	56.01	Un poco difícil	5.0
“¿Qué es la artritis reumatoide?”	57.31	Normal	53.26	Un poco difícil	5.8
“Colecistitis”	62.01	Normal	57.26	Un poco difícil	5.2
“Dieta para pacientes con Enfermedad inflamatoria intestinal”	66.16	Bastante fácil	64.55	Adecuado	5.1
“Tu hijo quiere una mascota: analiza los pros y los contras”	60.91	Normal	55.24	Un poco difícil	5.5
“Consejos para el niño que lleva tupper a la escuela”	65.85	Bastante fácil	60.68	Un poco difícil	5.1

Tabla 14 – Resultados Legible.es 2ª parte. Elaboración propia.

*No es propiamente un índice sino una nueva escala para el índice Szigriszt-Pazos que redefine la escala original propuesta por Szigriszt, ya que, según su autora Inés Mª Barrio Cantalejo, la escala precisaba de una adaptación al haber sido obtenida con una muestra insuficiente, no representativa ni aleatoria de textos.

Cabría esperar que el resultado que proporciona la herramienta de *legible.es* tanto para el índice Fernández-Huerta como para el Szigriszt-Pazos (Flesch-Szigriszt) fuera el mismo que el proporcionado por INFLESZ ya que el programa aplica la misma fórmula matemática y el texto utilizado ha sido el mismo en ambos casos, sin embargo, la puntuación obtenida para ambas fórmulas difiere ligeramente. Esto es debido a que estas fórmulas dependen de variables como el total de sílabas, el número de palabras o de frases y cada software

utiliza un algoritmo distinto para contabilizarlas, es decir, para un mismo texto un software puede contabilizar un número de frases, de sílabas o de palabras diferente produciendo discordancias en el resultado obtenido.

5.1.4. Legibilidadmu.cl

Se ha calculado la legibilidad lingüística de las muestras utilizando el índice µ, mediante la herramienta online *legibilidadmu.cl*. En la *tabla 15* se encuentran los resultados obtenidos.

Artículo	Índice µ	Dificultad
"La artritis reumatoide"	44.2	Difícil
"Vía biliar: Cólico biliar, colecistitis, colangitis y coledocolitiasis ¿Cómo distinguirlas?"	39.2	Difícil
"Guía del paciente a los marcadores tumorales"	45	Difícil
"Artritis idiopática juvenil"	44.9	Difícil
"Enfermedad por el virus de Zika"	47.4	Difícil
"Chikungunya"	45.6	Difícil
"Artritis reumatoide"	46.4	Difícil
"Colecistitis aguda"	58.7	Un poco difícil
"Decálogo de actuación en los colegios ante las alergias"	49.5	Difícil
"La hipoglucemia"	50.5	Difícil
"¿Cómo se evalúa el dolor en los niños?"	54.5	Un poco difícil
"¿Qué es la artritis reumatoide?"	53.3	Un poco difícil
"Colecistitis"	57.4	Un poco difícil
"Dieta para pacientes con Enfermedad inflamatoria intestinal"	63	Adecuado
"Tu hijo quiere una mascota: analiza los pros y los contras"	55.2	Un poco difícil
"Consejos para el niño que lleva tupper a la escuela"	60.8	Un poco difícil

Tabla 15 – Resultados Legibilidadmu.cl. Elaboración propia.

5.2. Análisis de los resultados proporcionados por cada herramienta

Los textos de muestra proporcionados únicamente indican un valor de complejidad en una escala de 3 niveles, es decir, no se dispone de un valor numérico que se pueda comparar con el proporcionado por las herramientas. Lo que se ha hecho es adaptar cada escala, normalmente compuesta por 7 o 5 niveles, a una

escala propia, de modo que los resultados devueltos por cada herramienta adapten lo mejor posible a los niveles de dificultad reales de cada texto. No se puede asegurar la validez de las escalas de 3 niveles que se han definido para cada índice y herramienta, puesto que no se dispone de una muestra de textos suficientemente grande ni representativa de la población.

5.2.1. INFLESZ

En la *figura 7* se observa la gráfica con los valores del índice Fernández-Huerta, en azul, y la recta de regresión. Se ha representado el valor del índice recorriendo la *tabla 10* en sentido descendente, del grupo difícil al fácil (pero dentro de cada agrupación de textos estos no están ordenados siguiendo ningún patrón). La recta de regresión tiene pendiente positiva, como cabía esperar, puesto que en el índice Fernández-Huerta valores más pequeños indican menor legibilidad mientras que valores más grandes indican mayor legibilidad.
Lo mismo ocurre con los valores del índice Flesch-Szigriszt, representado gráficamente en la *figura 8*, también con recta de regresión positiva por la misma razón que en el caso anterior.

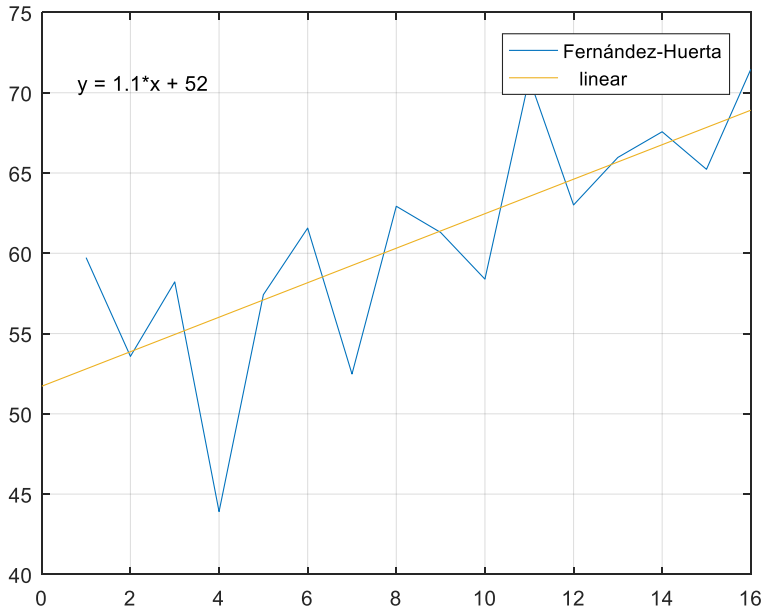

Figura 7 – Gráfica Índice Fernández-Huerta con recta de regresión.

Si se observa la *figura 9*, resulta trivial ver que existe una correlación entre los valores del índice Fernández-Huerta y Flesch-Szigriszt, puesto que la forma de ambas gráficas es muy similar. Utilizando Matlab, se ha obtenido el coeficiente de correlación de Pearson, el cual puede utilizarse para medir cual es el grado de relación entre dos variables cuantitativas.

$$r = 0.9996$$
$$p = 6.7976 \cdot 10^{-23}$$

El valor de correlación r, muy cercano a 1, indica una correlación positiva y muy fuerte entre ambas variables. Además, como el p-valor presenta un valor inferior 0.05, se puede afirmar que las dos variables están relacionadas.

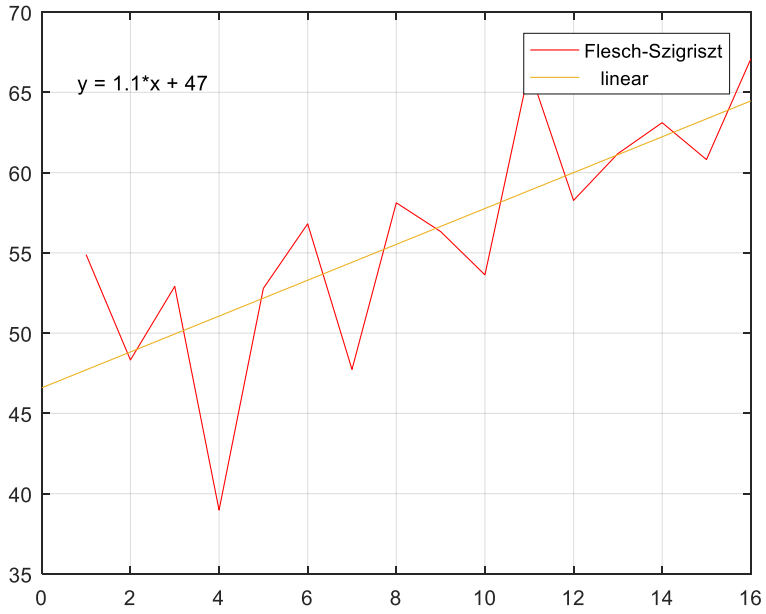

Figura 8 - Gráfica Índice Flesch-Szigriszt con recta de regresión.

Figura 9 – Gráfica Índices Fernández-Huerta y Flesch-Szigriszt

A continuación, se han agrupado los textos según su dificultad y se ha realizado un estudio de la media y la desviación típica de cada grupo, tal y como se ha explicado en el *apartado 4.2*. De este modo se ha obtenido:

Escala de 3 niveles **Fernández-Huerta (INFLESZ)**:

Difícil	Media	Fácil
0 – 58.1	58.1 – 63.8	63.8 – 100

El porcentaje de acierto utilizando esta escala es del **75%** respecto a los textos de muestra.

Escala de 3 niveles **Flesch-Szigriszt (INFLESZ)**:

Difícil	Media	Fácil
0 – 53.1846	53.1846 – 59.0217	59.0217 – 100

El porcentaje de acierto utilizando esta escala es del **81.25%** respecto a los textos de muestra

5.2.2. Professional Spanish Lexile Analyzer

En el caso de la escala Lexile los valores se comportan al revés que en el resto de índices que se están analizando, es decir, valores altos significan textos de legibilidad difícil y valores bajos de legibilidad fácil, es por esta razón que la recta de regresión tiene pendiente negativa (*figura 10*).

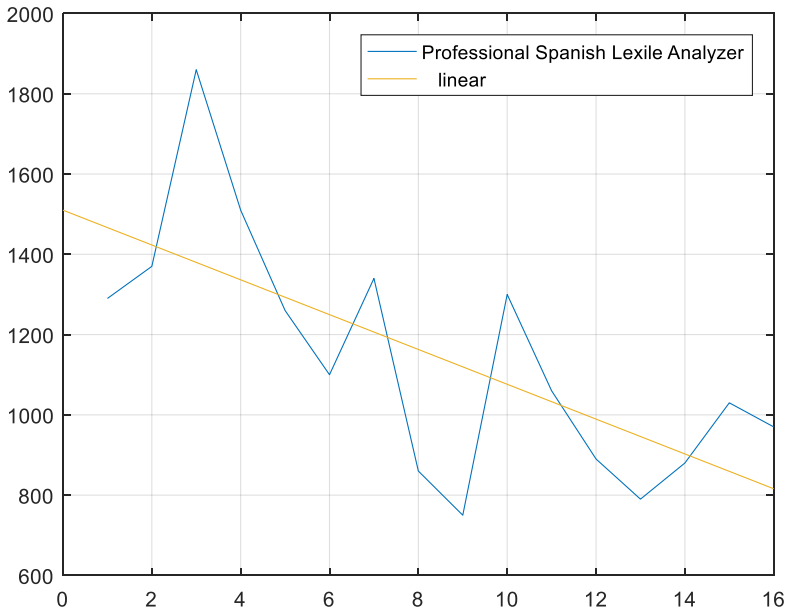

Figura 10 – Gráfica resultados Professional Spanish Lexile Analyzer con recta de regresión.

Escala de 3 niveles **Professional Lexile Analyzer**:

Fácil	Media	Fácil
0L (o inferior) – 924L	924L – 1272.L	1272L – 1700L (o superior)

El porcentaje de acierto utilizando esta escala es del **50%** respecto a los textos de muestra. En este caso el porcentaje obtenido es bastante más bajo, debido a que los valores obtenidos dentro de cada grupo son más variantes.

5.2.3. legibilidadmu.cl

En este caso, también se aprecia una tendencia de crecimiento en la gráfica de la figura 11, puesto que el valor de la legibilidad μ es más bajo a mayor complejidad del texto.

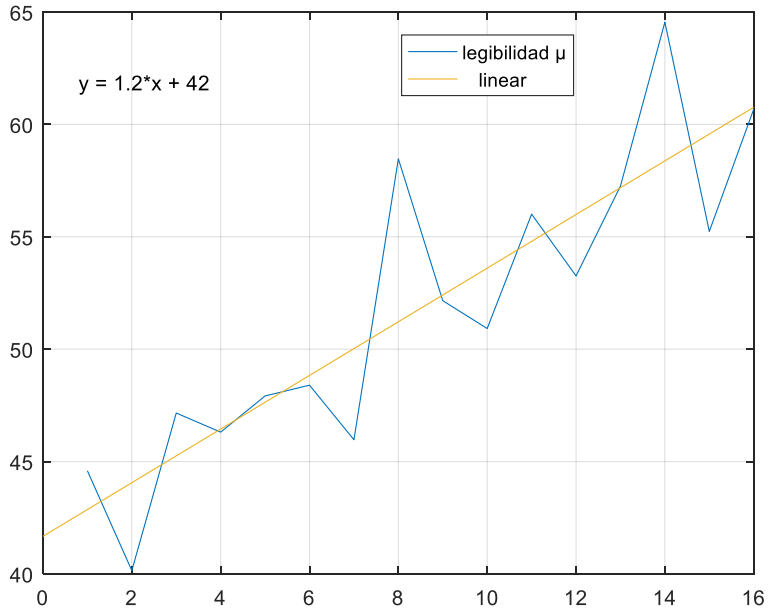

Figura 11 – Gráfica de resultados para legibilidad µ con recta de regresión.

Escala de 3 niveles **legibilidad µ (legibilidadmu.cl)**:

Difícil	Media	Fácil
0 – 47.4	47.4 – 54.8	54.8 – 100

El porcentaje de acierto utilizando esta escala es del **75%** respecto a los textos de muestra.

5.2.4. Legible.es

Se ha obtenido el coeficiente de correlación de Pearson, uno a uno para cada índice que calcula la herramienta online de *legible.es (tabla 16).* Tanto la gráfica de valores del índice INFLESZ como el estudio de su correlación con el resto de índices se corresponden exactamente con los del índice Szigriszt-Pazos, pues como ya se ha comentado, el valor numérico es exactamente el mismo, lo único que cambia es como se interpreta este valor numérico a la hora de asignar un valor de complejidad en la escala. El valor de los p-valores se ha omitido porque en todos los casos es muy inferior a 0.05.

	Fernández-Huerta	**Gutiérrez**	**Szigriszt-Pazos**	**Legibilidad µ**
Fernández-Huerta	1	0.9826	0.9993	0.8014
Gutiérrez	0.9826	1	0.9828	0.8113
Szigriszt-Pazos	0.9993	0.9828	1	0.8106
Legibilidad µ	0.8014	0.8113	0.8106	1

Tabla 16 – Correlación entre los índices que calcula la herramienta de legible.es.

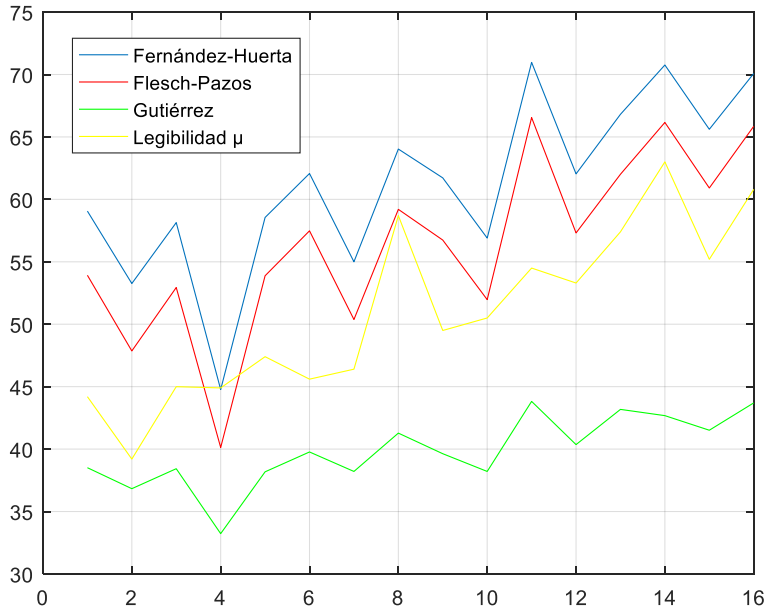

Figura 12 – Gráfica con los cuatro[23] índices de legible.es

En la *figura 16* se aprecia que todos los índices tienen una tendencia a crecer, tal y como cabría esperar, por la naturaleza de las fórmulas (valores bajos para textos difíciles y altos para textos fáciles) y la gran correlación que existe entre ellas (*tabla 16*).

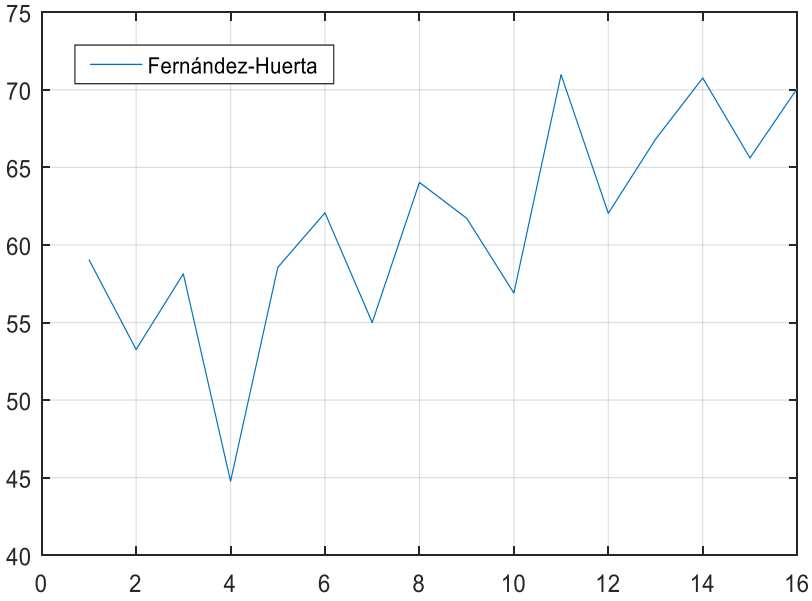

Figura 13 – Gráfica del índice Hernández-Huerta de legible.es

Escala de 3 niveles **Fernández-Huerta (legible.es)**:

Difícil	Media	Fácil
0 – 58.47	58.47 – 63.96	63.96 – 100

[23] El índice INFLESZ que devuelve *legible.es* tiene el mismo valor numérico que el Szigriszt-Pazos, por eso no se ha representado.

El porcentaje de acierto utilizando esta escala es del **62.5%** respecto a los textos de muestra.

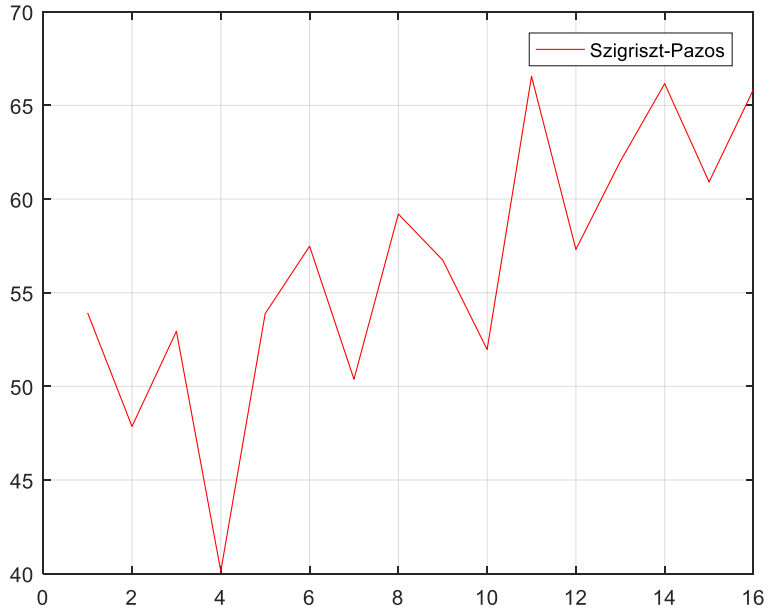

Figura 14 – Gráfica del índice Szigriszt-Pazos de legible.es

Escala de 3 niveles **Szigriszt-Pazos (legible.es)**:

Difícil	Media	Fácil
0 – 53.52	53.52 – 59.2	59.2 – 100

El porcentaje de acierto utilizando esta escala es del **62.5%** respecto a los textos de muestra.

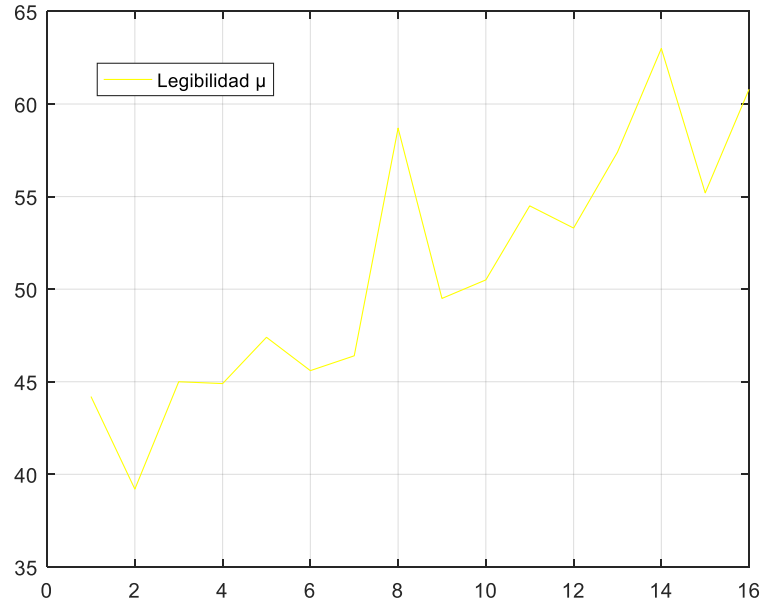

Figura 15 – Gráfica de legibilidad μ de legible.es

Escala de 3 niveles **legibilidad μ (legible.es)**:

Difícil	Media	Fácil
0 – 46	46 – 54.4	54.4 – 100

El porcentaje de acierto utilizando esta escala es del **75%** respecto a los textos de muestra.

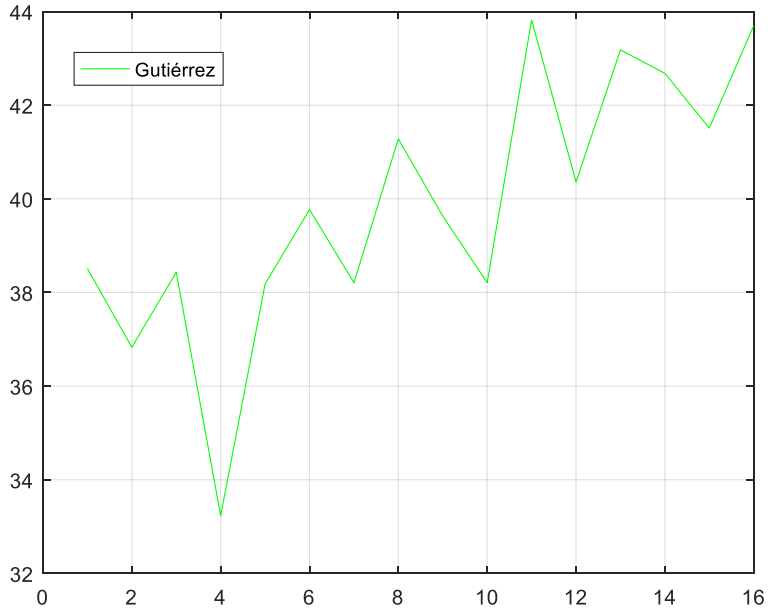

Figura 16 – Gráfica del índice Gutiérrez de legible.es

Escala de 3 niveles **Gutiérrez (legible.es)**:

Difícil	Media	Fácil
0 – 38.7	38.7 – 40.9	40.9 – 100

El porcentaje de acierto utilizando esta escala es del **75%** respecto a los textos de muestra.

Con el método que se ha utilizado para determinar las escalas de interpretación, el mejor resultado ha sido el proporcionado por el programa INFLESZv1.0 utilizando la fórmula de Flesch-Szigriszt, el cual ha obtenido un porcentaje de acierto del 81.25%. En cambio, para la misma fórmula, pero aplicando el algoritmo de la herramienta de *legible.es* el porcentaje de acierto ha sido del 62.5%, y aunque los valores varían ligeramente debido a algoritmos internos a la hora de realizar el conteo de palabras, frases y sílabas, la variación en la puntuación para cada muestra no dista mucho de la obtenida con INFLESZv1.0.
Si se realiza un ligero ajuste sobre los valores límite de la escala obtenidos por el método de la media y la desviación típica, se puede llegar a obtener el mismo porcentaje de acierto que con INFLESZv1.0.

Escala de 3 niveles **Szigriszt-Pazos (legible.es)**:

Difícil	Media	Fácil
0 – **55**	**55 – 60**	60 – 100

Tabla 17 – Escala de 3 niveles para interpretar el resultado del índice Szigriszt-Pazos

Con esta pequeña modificación el porcentaje de ajuste ha pasado de ser del 62.5% al **81.25%**, el cual representa una mejora sustancial.
Por tanto, **se va a utilizar el índice de Szigriszt-Pazos con la escala de 3 niveles que se ha definido en la *tabla 17.*** Una vez más, cabe destacar que la efectividad de esta escala no está comprobada estadísticamente debido al tamaño muestral utilizado.

5.3. Desarrollo del software

Para el desarrollo del software se ha dividido en diseño en bloques funcionales (*Figura 10*). A continuación, se va a explicar cada uno de esos bloques.
Primer bloque ("Inicio"): se encarga de realizar el procesado de los argumentos de la línea de comandos. Para ello se hace uso de la librería *argparse* (estándar de Python).

```
parser = argparse.ArgumentParser()
parser.add_argument("url", help="URL de la web a analizar. De la forma
http://www.example.com")
args = parser.parse_args()
```

Únicamente se ha añadido como parámetro de entrada la URL de la web que se desea analizar.

Segundo bloque ("Bloque de petición HTTP"): es el encargado de realizar la petición al servidor HTTP para recuperar la información de la web indicada en la llamada al programa. Está dividido en dos bloques:

- Genera un objeto Request, a partir de la URL de entrada y permite configurar las cabeceras de la petición HTTP.

```
req = urllib.request.Request(
    url=args.url,
    data=None,
    headers={
        'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64)' #Algunas webs
pueden bloquear el acceso si el User-Agent no se cambia.
    }
)
```

- Realiza la petición del recurso al servidor HTTP.

```
html = urllib.request.urlopen(req).read() #Se hace el request de la web al
servidor HTTP y se guarda el HTML en la variable html
```

Tercer bloque ("Bloque de procesado HTML"): es el bloque encargador de procesar el fichero HTML que se ha recibido. Está dividido en cuatro bloques:

- Realiza una simplificación del contenido HTML recibido, eliminando toda la información innecesaria, como los menús, pies de página, etc. Dejando solo el contenido principal de la web, donde se encuentra el cuerpo del artículo. Para ello hace uso de la librería *readability-lxml*.

```
readable_article = Document(html).summary() #Se sintetiza el HTML, eliminando
informacion no relevante para el analisis (menu,footer,etc.)
```

- Utilizando la librería *BeautifulSoup*, parsea la web simplificada eliminando etiquetas HTML que no resulten de interés, como el código Javascript o información de estilo que no tiene que ver con el contenido.

```
#Eliminamos las etiquetas <script> y <style>
for script in soup(["script", "style"]):
    script.extract()
```

- Mediante la función *prettify()* de la librería *BeautifulSoup*, reorganiza el código HTML para evitar posibles errores como etiquetas sin cerrar o cerradas incorrectamente.

```
html2=soup.prettify()
```

- Finalmente extrae el contenido en texto de la página, eliminando todas las etiquetas HTML y obteniendo un fichero de output con el cuerpo del artículo.

```
soup2 = BeautifulSoup(html2, 'html.parser')
texto = soup2.get_text()
```

Cuarto bloque ("Calcula legibilidad lingüística"): Accede a ese fichero y realiza el cálculo de la legibilidad lingüística del texto, devolviendo un valor numérico que es interpretado según la escala que se ha definido previamente.

```
indice = legibilidad.szigriszt_pazos(texto)
print("El indice S-P es: ",indice)
print("El valor en mi Escala es: ", legibilidad.miEscala(indice))
```

Quinto bloque ("Genera XML"): Genera un documento .XML que almacena la URL de la web, el título del artículo, el cuerpo del artículo y el nivel de dificultad asignado según la escala de 3 niveles.

```
soupXml = BeautifulSoup(features='xml')
tag_url=soup.new_tag("url")
tag_url.string=args.url
tag_titulo=soup.new_tag("titulo")
tag_titulo.string=readable_title
tag_contenido=soup.new_tag("contenido")
tag_contenido.string=texto
tag_legibilidad=soup.new_tag("legibilidad")
tag_legibilidad.string=legibilidad.miEscala(indice)
soupXml.append(tag_url)
soupXml.append(tag_titulo)
soupXml.append(tag_contenido)
soupXml.append(tag_legibilidad)
```

Finalmente, se guarda este fichero XML con nombre igual a la fecha y hora de la consulta (para evitar que el título se repita en consultas futuras).

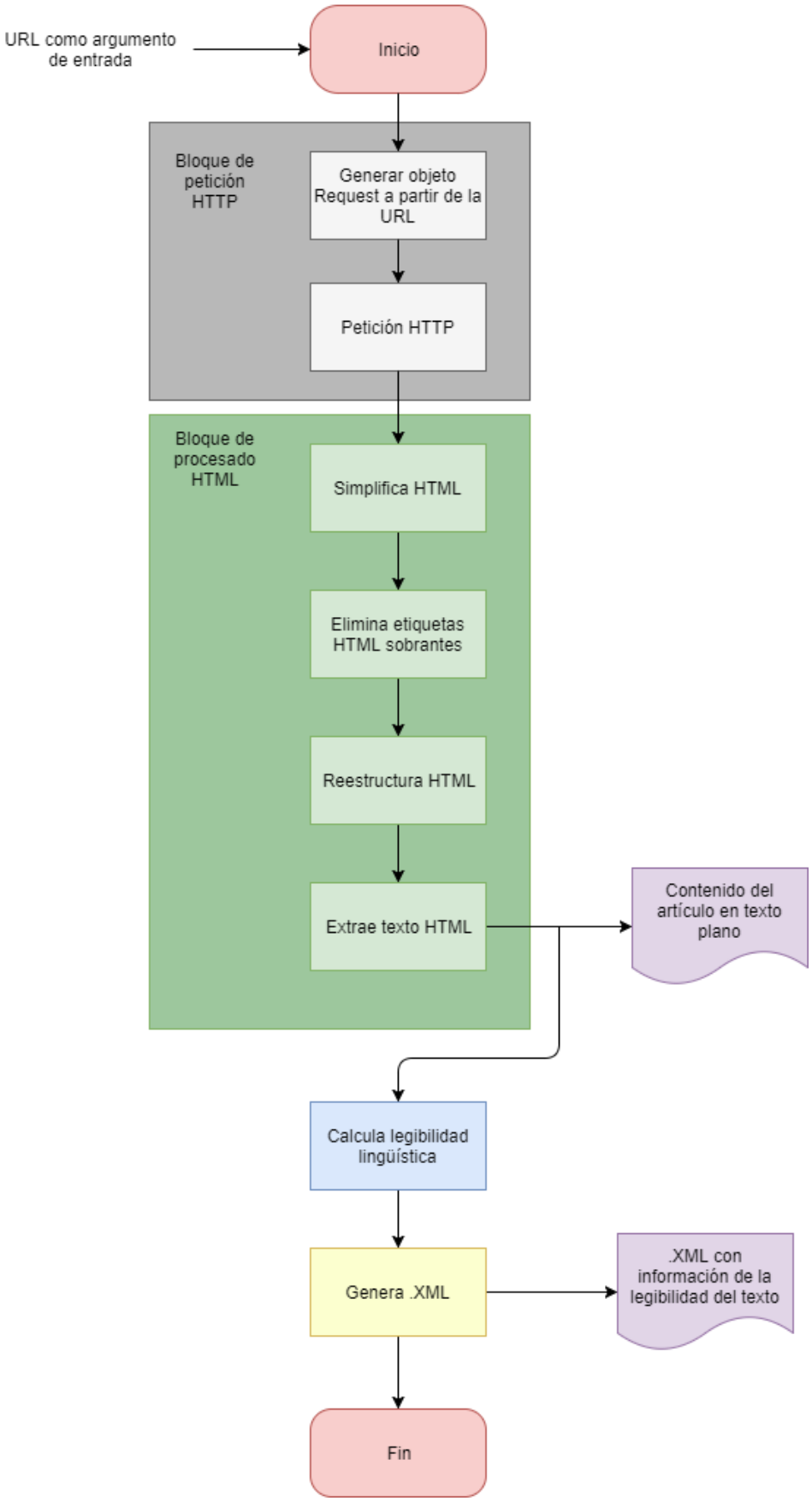

Figura 17 – Diagrama de flujo de la aplicación.

5.4. Dificultad de los artículos obtenidos de Google

Para evaluar la dificultad de los artículos que Google muestra en primeras posiciones al realizar una búsqueda sobre cualquier tema de salud, se ha utilizado la herramienta que se ha desarrollado. Se han elaborado 7 tablas que contienen 10 artículos cada una. **Se ha calculado la legibilidad de cada artículo utilizando el programa que se ha desarrollado.** Para indicar la legibilidad que el programa ha asignado a cada artículo se ha utilizado el **mismo código de colores que se utilizó para las muestras de texto proporcionadas**.

Cáncer	
Título del artículo	**Dificultad asignada por el programa**
"¿Qué es el cáncer?" https://www.cancer.org/es/cancer/aspectos-basicos-sobre-el-cancer/que-es-el-cancer.html	Media
"¿Qué es el cáncer y cómo se desarrolla?" http://www.seom.org/en/informacion-sobre-el-cancer/que-es-el-cancer-y-como-se-desarrolla	Media
"Cáncer de mama: Síntomas y signos" http://www.cancer.net/es/tipos-de-c%C3%A1ncer/c%C3%A1ncer-de-mama/s%C3%ADntomas-y-signos	Fácil
"Signos y síntomas del cáncer" https://www.cancer.org/es/cancer/aspectos-basicos-sobre-el-cancer/senales-y-sintomas-del-cancer.html	Fácil
"10 síntomas de cáncer que muchos ignoran" https://mejorconsalud.com/10-sintomas-cancer-que-muchos-ignoran/	Fácil
"10 síntomas del cáncer que pueden pasar desapercibidos" http://www.bbc.com/mundo/noticias/2016/04/160429_salud_cancer_10_sintomas_generales_il	Fácil
"Tratamientos para el cáncer" https://medlineplus.gov/spanish/ency/patientinstructions/000901.htm	Difícil
"Tipos de tratamiento" https://www.cancer.gov/espanol/cancer/tratamiento/tipos	Fácil
"Cáncer: MedlinePlus enciclopedia médica" https://medlineplus.gov/spanish/ency/article/001289.htm	Fácil
"¿Qué es el cáncer? - Instituto Nacional del Cáncer" https://www.cancer.gov/espanol/cancer/naturaleza/que-es	Media

Tabla 18 – Dificultad de los artículos con palabra clave "Cáncer". Elaboración propia.

Lupus	
Título del artículo	**Dificultad asignada por el programa**
"Lupus | Síntomas del Lupus" https://medlineplus.gov/spanish/lupus.html	Fácil
"¿Qué es el lupus?" https://www.niams.nih.gov/Portal_en_espanol/Informacion_de_salud/Lupus/default.asp	Fácil
"Lupus: tratamientos, síntomas e información" http://www.cuidateplus.com/enfermedades/musculos-y-huesos/lupus.html	Difícil
"Preguntas y Respuestas sobre el Lupus"	Media

http://www.felupus.org/preguntas.php	
“Diagnóstico del lupus” http://www.webconsultas.com/lupus/diagnostico-del-lupus-1963	Media
“Tratamiento del Lupus” https://www.adeluna.es/que-es-lupus/tratamiento-del-lupus/	Difícil
“¿Qué es el Lupus?” http://www.med.ub.es/MIMMUN/UCMAS/CASTELLA/INFCAST.HTM	Fácil
“Lupus: una enfermedad sin cura” http://www.cuidateplus.com/enfermedades/musculos-y-huesos/lupus/2015/05/10/lupus-enfermedad-cura-69763.html	Difícil
“Lupus tratamientos” https://www.news-medical.net/health/Lupus-Treatments-(Spanish).aspx	Difícil
“Como vivir con Lupus” http://www.lupusny.org/about-lupus/espanol/c%C3%B3mo-vivir-con-lupus	Fácil

Tabla 19 – Dificultad de los artículos con palabra clave “Lupus”. Elaboración propia.

Gripe (Influenza)	
Título del artículo	**Dificultad asignada por el programa**
“¿Que es la gripe? Cuáles son sus síntomas y tratamiento” http://blog.plandecalidadsns.es/que-es-la-gripe-cuales-son-sus-sintomas-y-tratamiento/	Fácil
“OMS | Gripe” http://www.who.int/topics/influenza/es/	Difícil
“Tratamiento de la gripe” http://www.webconsultas.com/gripe/tratamiento-de-la-gripe-327	Difícil
“Gripe, qué es y tipos de gripe” http://www.webconsultas.com/gripe/gripe-323	Fácil
“Gripe tratamientos, síntomas e información en CuidatePlus” http://www.cuidateplus.com/enfermedades/infecciosas/gripe.html	Fácil
“Gripe: consejos de prevención y tratamiento” https://www.ocu.org/salud/enfermedades/noticias/gripe-consejos-de-prevencion-y-tratamiento531204	Fácil
“¿Qué tomo para combatir la gripe?” https://www.ocu.org/salud/medicamentos/noticias/gripe-que-tomar-y-que-no-tomar-561014	Media
“Síntomas de la influenza y sus complicaciones” https://espanol.cdc.gov/enes/flu/about/disease/complications.htm	Media
“Previniendo la gripe” https://es.familydoctor.org/previniendo-la-gripe/	Fácil
“Causas de la gripe” http://www.webconsultas.com/gripe/causas-de-la-gripe-324	Difícil

Tabla 20 – Dificultad de los artículos con palabra clave “Gripe”. Elaboración propia.

Diabetes	
Título del artículo	**Dificultad asignada por el programa**

“Diabetes tratamientos, síntomas e información en CuidatePlus” http://www.cuidateplus.com/enfermedades/digestivas/diabetes.html	Media
“¿Cuáles son los diferentes tipos de diabetes?” http://www.redgdps.org/?idregistro=296	Fácil
“Tipos de diabetes” http://www.fundaciondiabetes.org/infantil/177/tipos-de-diabetes-ninos	Media
“Diabetes tipo 2: ¿Cuál es el tratamiento?” http://kidshealth.org/es/teens/treat-type2-esp.html	Media
“Tratamiento de la diabetes tipo 2” https://dtc.ucsf.edu/es/tipos-de-diabetes/diabetes-tipo-2/tratamiento-de-la-diabetes-tipo-2/	Difícil
“Diabetes: MedlinePlus enciclopedia médica” https://medlineplus.gov/spanish/ency/article/001214.htm	Difícil
“Qué es la diabetes” http://www.fundaciondiabetes.org/prevencion/309/que-es-la-diabetes-2	Fácil
“Qué es la diabetes” https://www.cdc.gov/diabetes/spanish/basics/diabetes.html	Difícil
“Qué es la diabetes” http://www.clubdelhipertenso.es/index.php/diabetes	Media
“7 síntomas que podrían indicarte que sufres de diabetes” http://descubretusalud.com/7-sintomas-indicarte-sufres-diabetes/	Fácil

Tabla 21 – Dificultad de los artículos con palabra clave “Diabetes”. Elaboración propia.

Herpes	
Título del artículo	**Dificultad asignada por el programa**
“Herpes: Síntomas, causas y tratamiento” https://es.familydoctor.org/condicion/herpes/	Fácil
“¿Cuáles son los Signos y Síntomas del Herpes?” http://articulos.mercola.com/herpes/sintomas.aspx	Media
“Enfermedades de transmisión sexual” https://www.cdc.gov/std/spanish/herpes/stdfact-herpes-s.htm	Fácil
“Herpes genital – La infección por el virus del herpes simple” https://www.medicina21.com/Articulos/V75/Herpes-genital-La-infeccion-por-el-virus-del-herpes-simple.html	Media
“Infecciones por el virus del herpes” http://www.revistasbolivianas.org.bo/scielo.php?pid=S2304-37682010001000007&script=sci_arttext	Difícil
“Herpes simple. Herpes labial” http://netdoctor.elespanol.com/articulo/herpes-simple	Fácil
“Tratamiento del herpes” http://www.webconsultas.com/herpes/tratamiento-del-herpes-337	Media
“Tratamiento del Herpes Simple” https://www.news-medical.net/health/Herpes-Simplex-Treatment-(Spanish).aspx	Fácil
“Virus del herpes simple” http://www.who.int/mediacentre/factsheets/fs400/es/	Fácil
“Herpes genital Tratamiento” http://www.onmeda.es/enfermedades/herpes_genital-tratamiento-1850-6.html	Media

Tabla 22 - Dificultad de los artículos con palabra clave "Herpes". Elaboración propia.

Alzheimer	
Título del artículo	**Dificultad asignada por el programa**
"Las diez señales" http://www.alz.org/espanol/signs_and_symptoms/las_10_senales.asp	Fácil
"Fases del Alzheimer" https://knowalzheimer.com/todo-sobre-el-alzheimer/fases-del-alzheimer/	Difícil
"Alzheimer tratamientos, síntomas e información - CuidatePlus" http://www.cuidateplus.com/enfermedades/neurologicas/alzheimer.html	Difícil
"Qué es el Alzheimer | enfermedades degenerativas" http://www.infosalus.com/enfermedades/neurologia/alzheimer/que-es-alzheimer-60.html	Difícil
"Entender la enfermedad de Alzheimer: lo que usted necesita saber" https://www.nia.nih.gov/health/entender-enfermedad-alzheimer-lo-usted-necesita-saber	Fácil
"Los signos y síntomas de la enfermedad de Alzheimer" https://www.brightfocus.org/espanol/la-enfermedad-de-alzheimer-y-la-demencia/enfermedad-de-alzheimer-sintomas-y-etapas	Difícil
"Cómo detectar primeros síntomas de Alzheimer, 10 signos" http://www.eltallerdemismemorias.com/2013/12/20/primeros-sintomas-de-alzheimer/	Fácil
"10 SÍNTOMAS TÍPICOS DE LA ENFERMEDAD DE ALZHEIMER" https://www.mdsaude.com/es/2016/09/sintomas-de-la-enfermedad-de-alzheimer.html	Media
"Enfermedad de Alzheimer" https://es.familydoctor.org/condicion/enfermedad-de-alzheimer/	Media
"Tratamientos de la Enfermedad de Alzheimer" https://www.brightfocus.org/espanol/la-enfermedad-de-alzheimer-y-la-demencia/tratamientos-de-la-enfermedad-de-alzheimer	Difícil

Tabla 23 - Dificultad de los artículos con palabra clave "Alzheimer". Elaboración propia.

SIDA (VIH)	
Título del artículo	**Dificultad asignada por el programa**
"Sida tratamientos, síntomas e información en CuidatePlus" http://www.cuidateplus.com/enfermedades/infecciosas/sida.html	Media
"¿Qué es el VIH? | infoSIDA" http://www.infosida.es/que-es-el-vih	Difícil
"Cuáles son los Síntomas & Signos del VIH/SIDA?" https://www.plannedparenthood.org/es/temas-de-salud/enfermedades-de-transmision-sexual-ets/vih-sida/cuales-son-los-sintomas-del-vihsida	Fácil
"Tratamiento para la infección por el VIH: Conceptos básicos" https://infosida.nih.gov/understanding-hiv-aids/fact-sheets/21/51/tratamiento-para-la-infeccion-por-el-vih--conceptos-basicos	Fácil

“Sida tratamientos, síntomas e información - CuidatePlus” http://www.cuidateplus.com/enfermedades/infecciosas/sida.html	Media
“¿Qué es el SIDA?” http://www.aidsinfonet.org/fact_sheets/view/101?lang=spa	Fácil
“¿Qué es el SIDA?” https://www.aciprensa.com/sida/definicion.htm	Difícil
“Tratamiento contra el VIH y el sida” http://www.uncares.org/es/content/tratamiento-contra-el-vih-y-el-sida	Media
“¿Qué tratamientos existen para el VIH/Sida? ¿Por qué es tan importante tomarlos correctamente?” https://apoyopositivo.org/faq/tratamientos/tratamientos-vih-sida/	Difícil
“Tratamiento del sida” http://www.webconsultas.com/sida/tratamiento-del-sida-367	Media

Tabla 24 - Dificultad de los artículos con palabra clave “Sida”. Elaboración propia.

	Fácil	Media	Difícil
Cáncer	6	3	1
Lupus	4	2	4
Gripe	5	2	3
Diabetes	3	4	3
Herpes	5	4	1
Alzheimer	3	2	5
Sida	3	4	3

Tabla 25 – Recuento de textos clasificados en dificultad.

De estos resultados, podemos obtener la media de dificultad de los artículos proporcionados por el buscador de Google:

- El **41.43%** de textos son **fáciles**
- El **30%** de textos son de **dificultad media**
- El **28.57%** de textos son **difíciles**

6. Discusión

Las herramientas analizadas proporcionan una indicación de la legibilidad lingüística bastante acertada, teniendo en cuenta los límites que presenta cualquier herramienta que base su funcionamiento en el uso de estas fórmulas, las cuales no deben ser tomadas como una fuente infalible de información de legibilidad.
Sin embargo, al analizar los resultados obtenidos para cada muestra, se ha podido comprobar que la diferencia de puntuación para los textos de tipo fácil, medio y difícil es bastante pequeña para los textos de muestra proporcionados, dificultando la tarea de establecer una escala de tres niveles correctamente acotada. Si el número de textos de muestra hubiese sido más grande, los límites hubiesen quedado mejor definidos y probablemente se hubiesen obtenido escalas más adecuadas para este tipo de textos.
La herramienta que se ha desarrollado parece funcionar bien, no obstante, para validar su funcionamiento, hubiese sido conveniente una vez más disponer de una muestra de textos, ya catalogados, y sobre los que no se hubiese trabajado previamente, para comprobar que porcentaje de ellos eran correctamente clasificados con el programa que se ha desarrollado.
Respecto a los resultados de Google al realizar las búsquedas con las palabras clave más comunes, en algunos casos la proporción entre textos fáciles y los medios y difíciles podría ser tolerable, como en el caso del cáncer y el herpes, sin embargo, en el resto de casos, los resultados devueltos deberían ser revisados con el objetivo de proporcionar al usuario información fácilmente comprensible, independientemente de su perfil.
Resulta interesante que, tal y como se ha comentado en el párrafo anterior, mientras enfermedades como el cáncer y el herpes están documentadas con una serie de artículos de dificultad media asequible, la gripe, que resulta bastante más común, cuenta con una cantidad de textos difíciles bastante grande, cuando, debido a su recurrencia, debería aparecer mucha información asequible para cualquier usuario al realizar una búsqueda. El valor medio de textos fáciles devuelto por Google en la prueba que se ha realizado es del 41.43%, es decir, en media, menos de la mitad de los artículos que Google nos ha sugerido en las primeras posiciones tienen un nivel de complejidad asequible para cualquier persona. Por otro lado, los textos de dificultad media representan el 30% y los textos difíciles el 28.57% restante, esto supone una mayoría de textos cuya complejidad es probablemente demasiado exigente para la mayoría de pacientes y usuarios de Internet sin conocimientos avanzados de salud. Es especialmente importante que la información proporcionada por buscadores no específicos sea comprensible para todos, es decir, aumentar la cantidad de resultados cuya dificultad de comprensión sea fácil, dando prioridad a estos frente a publicaciones más complejas que deberían ser accesibles desde buscadores especializados (los cuales también deberían facilitar información de fácil legibilidad a sus usuarios, si su perfil así lo requiere).

7. Conclusiones y futuras líneas de trabajo

Los objetivos del trabajo eran el desarrollo de un software capaz de establecer un nivel de dificultad en la comprensión de un texto de cualquier ámbito del campo de la salud y la utilización de dicho software para evaluar los resultados devueltos por un buscador de uso general (en este caso Google por ser el más usado) para un conjunto de palabras clave. Para satisfacer estos objetivos se plantearon una serie de hitos y objetivos intermedios. El primero era la realización de una investigación del estado del arte, la cual ha permitido conocer las fórmulas de legibilidad existentes, los algoritmos y herramientas para llevarlo a cabo e información de cómo el usuario de internet consulta información médica online.
A continuación, se han probado las diferentes herramientas sobre el conjunto de muestras y se ha realizado un análisis sobre el resultado proporcionado, con el objetivo de conocer cuál es el algoritmo que mejor funciona para textos médicos y determinar así cuál ha de ser implementado en el software propio. En este aspecto, el resultado no parece concluyente, debido a la cantidad insuficiente de muestras disponibles para el análisis. Finalmente, se ha decidido utilizar la fórmula de Szigriszt-Pazos (también conocida como Flesch-Szigriszt) ya que se trata de una fórmula de efectividad comprobada y es la fórmula de referencia en la actualidad. Los resultados proporcionados por esta fórmula han sido interpretados utilizando la escala propia de 3 niveles elaborada en el *apartado 5.2*.
Una vez se ha determinado el algoritmo que se va a utilizar se ha desarrollado un programa que cumple con las especificaciones que se habían establecido, haciendo uso de diversas librerías de Python. De este modo **se ha podido cumplir el primer objetivo principal.**
A continuación, se ha realizado una búsqueda de las 7 palabras clave que más buscan los usuarios de Internet, para obtener 10 artículos para cada una de estas palabras, que han sido analizados con el software desarrollado. De este modo se ha podido determinar la cantidad media de textos de cada dificultad que Google sugiere cuando un usuario realiza una búsqueda de información médica, **cumpliendo el segundo objetivo principal.**
La herramienta desarrollada como resultado de este estudio ha permitido conocer el nivel medio de complejidad que presentan los resultados de los buscadores tradicionales y ha puesto de relieve la necesidad de estos buscadores de adaptarse a esta necesidad. La inclusión de técnicas para la medida de la legibilidad lingüística por parte de los buscadores resultaría beneficiosa para los pacientes y usuarios en general que demandan información sanitaria, especialmente hoy en día con la nueva relación que se establece entre el médico y su paciente, el cual adopta una actitud proactiva (con el nacimiento de la figura del paciente empoderado) y es partícipe, junto con su médico, de las decisiones acerca de su salud.
El software que se ha desarrollado tiene diversas aplicaciones potenciales, como por ejemplo la integración con buscadores tradicionales, no especializados en información médica, para filtrar la información descartando aquella que resulte de una complejidad innecesaria. Aunque el uso de algoritmos basados en fórmulas para el cálculo de la legibilidad puede resultar útil en muchos casos, no se debe tomar como una referencia infalible, ya que la legibilidad lingüística de un texto es una propiedad difícil de medir de forma objetiva. Estas fórmulas han estado sujetas a controversia desde su nacimiento en la lengua inglesa y son muchos los autores que se han planteado si son realmente útiles o incluso si pueden ser problemáticas (DuBay, 2004), sin embargo, a pesar de ser una técnica relativamente antigua (las primeras fórmulas en inglés datan de la década de 1920), las fórmulas se siguen utilizando hoy en día, con sus ventajas e inconvenientes.
El principal problema que presentan estas fórmulas es que basan su funcionamiento en el análisis de diversas variables objetivas, como el número de palabras, de frases o de sílabas, no obstante, existen multitud de factores que entran en juego para determinar la legibilidad de un texto, algunas de ellas difícilmente cuantificables. Por ejemplo, las que involucran al lector, como su conocimiento de la lengua de origen del texto o incluso su estado emocional u otras variables como el orden lógico en el que se exponen las ideas, el uso correcto del lenguaje o el uso de figuras retóricas.
En resumen, resulta complicado evaluar una propiedad como la legibilidad de un texto en base a variables objetivas cuando se trata de una característica que depende también de variables subjetivas y dependientes del receptor.

Respecto a las posibles ampliaciones y mejoras en cuanto a lo desarrollado en este proyecto, algunas de las posibles futuras líneas de trabajo serían:

- *Análisis de las herramientas existentes con una muestra de textos con información sanitaria más grande*, que permitiese un análisis estadístico más preciso.
- *Mejora de la detección del cuerpo del artículo en el fichero HTML*: Respecto al software desarrollado, actualmente funciona en la mayoría de páginas webs, sin embargo, en algunas webs la detección del texto puede ser parcial e ignorar partes que resultarían de interés para el análisis de la legibilidad lingüística del artículo. En este sentido una solución parcial podría ser el análisis manual de las webs que suelen producir una salida incompleta y realizar un parseo (análisis sintáctico) personalizado para dicha web cuando la petición HTTP vaya destinado a un dominio en concreto.
- *Implementación del software para navegadores*: navegadores como Firefox y Google Chrome soportan el uso de Plugins. Implementar un Plugin que calcule la legibilidad lingüística de las páginas webs que se buscan, realizando un filtrado que eliminase aquellos resultados de la búsqueda cuya complejidad sea excesiva. Además, indicaría la complejidad de la web a la que se ha accedido, evitando al usuario la pérdida de tiempo al conocer desde un primer momento si la información a la que ha accedido tiene un nivel de complejidad adecuado.

8. Bibliografía

Barca Fernández, I., Parejo Miguez, R., Gutiérrez Martín, P., Fernández Alarcón, F., Alejandre Lázaro, G., & López de Castro, F. (s. f.). La información al paciente y su participación en la toma de decisiones clínicas. *Atención Primaria*, 361-364.

Barrio Cantalejo, I. M. (2007). Legibilidad y salud: los métodos de medición de la legibilidad y su aplicación al diseño de folletos educativos sobre salud. Recuperado a partir de https://repositorio.uam.es/handle/10486/2488

Barrio-Cantalejo, I. M., Simón-Lorda, P., Melguizo, M., Escalona, I., Marijuán, M. I., & Hernando, P. (2008). Validación de la Escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes. En *Anales del Sistema Sanitario de Navarra* (Vol. 31, pp. 135–152). SciELO Espana. Recuperado a partir de http://scielo.isciii.es/scielo.php?pid=S1137-66272008000300004&script=sci_arttext&tlng=en

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283.

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, *27*(1), 11-28. https://doi.org/10.2307/1473169

DuBay, W. H. (2004). The Principles of Readability. *Online Submission*. Recuperado a partir de https://eric.ed.gov/?id=ED490073

Fernández Huerta, J. (1959). Medidas sencillas de lecturabilidad. *Consigna*, *214*, 29–32.

Flesch, R. (1948). A new readability yardstick. *The Journal of Applied Psychology*, *32*(3), 221-233.

Gala, N., Rapp, R., & Bel-Enguix, G. (2014). *Language Production, Cognition, and the Lexicon*. Springer.

Gunning, R. (1968). *Technique of Clear Writing* (Edición: Revised). New York, N.Y.: McGraw Hill Higher Education.

Kincaid, J. P., Fishburne, J., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel* (No. RBR-8-75). NAVAL TECHNICAL TRAINING COMMAND MILLINGTON TN RESEARCH BRANCH, NAVAL TECHNICAL TRAINING COMMAND MILLINGTON TN RESEARCH BRANCH. Recuperado a partir de http://www.dtic.mil/docs/citations/ADA006655

R, B., & Luz, M. (2005). Cambios en la relación médico-paciente y nivel de satisfacción de los médicos. *Revista médica de Chile*, *133*(1), 11-16. https://doi.org/10.4067/S0034-98872005000100002

Rodríguez Diéguez, J. L., Moro Berihuete, P., & Cabero Pérez, M. V. (1992). Ecuaciones de predicción de lecturabilidad. Recuperado a partir de https://gredos.usal.es/jspui/handle/10366/69423

Spaulding, S. (1956). A Spanish Readability Formula. *The Modern Language Journal*, *40*(8), 433-441. https://doi.org/10.2307/319744

Szigriszt Pazos, F. (2001). *Sistemas predictivos de legilibilidad del mensaje escrito : fórmula de perspicuidad* (info:eu-repo/semantics/doctoralThesis). Universidad Complutense de Madrid, Servicio de Publicaciones, Madrid. Recuperado a partir de http://eprints.ucm.es/tesis/19911996/S/3/S3019601.pdf

Tinker, M. A. (1963). *Legibility of print*. Iowa State University Press.