Software to Determine the Readability of Written Documents by Implementing a Variation of the Gunning Fog Index Using the Google Linguistic Corpus

Luis Carlos Rodríguez Timaná, Diego Fernando Saavedra Lozano, and Javier Ferney Castillo García(✉)

Grupo de Investigación en Electrónica Industrial y Ambiental – GIEIAM, Universidad Santiago de Cali, Cali, Colombia
javier.castillo00@usc.edu.co

Abstract. In English linguistics the Gunning Fog Index is used to determine the readability of texts. This methodology isn't as effective in the Spanish language because the complexity of words isn't determined by the number of syllables, unlike what happens in English. Therefore, a software was developed that allows us to estimate the readability of an academic text written in Spanish in a quantitative way. This software allows to compare the traditional methodology of the Gunning fog index and a modification to it, using the corpus linguistics for the Spanish language, based on thousands of texts digitized by Google, where the frequency of use of certain words is related. Texts produced by students from first to last semester were evaluated. Each text was subjected to the Gunning fog index assessment methodology and the corpus methodology, changing the percentage of complex words to the percentage of unknown words. In the evaluation of first semester texts it was found that the average fog index was 29.25, and an average of 37.9 complex words, for these same texts was found a modified fog index of 18.62 and 5.1 unknown words. On the other hand, for the evaluation of the texts produced in the last semester, the average fog index was 27.55 and an average of 51.4 complex words, with the modified fog index was an average of 15.08 and 7.1 unknown words. With this study, aspects related to the best use of punctuation marks and the increase of vocabulary related to the profession can be identified in a quantitative way.

Keywords: Corpus linguistics · Gunning · Fog Index

1 Introduction

Readability is the ease with which a reader can understand a piece of writing. In natural language, the readability of the text depends on its content (the complexity of its vocabulary and syntax) and its presentation (typographical aspects such as font size, line height and line length). The easier a text is to read, the more readable it will be. Readability depends on whether a text is composed of short sentences, if it uses structures that allow

M. Botto-Tobar et al. (Eds.): ICAT 2019, CCIS 1193, pp. 409–420, 2020.
https://doi.org/10.1007/978-3-030-42517-3_31

the reader to advance in the content of the text, to place the key words properly in the right place, to keep a logical order, among other characteristics [1].

The readability of texts is a matter of interest for educators, publishers, journalists and others who use written texts as a means of diffusion. On several occasions, all these people must make decisions about the material they are going to use or disseminate. Tasks that tend to be time-consuming for lack of judgment [2].

Currently there are different methodologies that allow quantifying the readability of a text, but these are oriented to the English language, so there is a need to implement a tool that allows a simple way to evaluate texts in Spanish. A software based on a variation of the Gunning fog index was developed for evaluate texts in Spanish, using Google Ngram by means of an API. The software was implemented in texts of 10 students of different semesters of the Faculty of Engineering of the Universidad Santiago de Cali. It was determined that first semester students had a lower readability in their texts than students in more advanced semesters.

Theoretical Framework. According to the dictionary of the "Real Academia Española", readability is the quality of being read. The easier a text is to read and understand, the more readable it will be, which is why it is a very important factor when creating content, especially when it is educational content [3]. There are different methodologies for quantitatively calculating the readability of texts. Some of them are mentioned below.

Flesch's readability test evaluates texts on a 100-point scale and considers the number of words per sentence and the average number of syllables per word. It does not consider any variable that may be affected by the language in which the text is written [3]. The formula for this test is mentioned in Eq. 1.

$$206.835 - 1.01 * \left(\frac{total\ words}{total\ sentences}\right) - 84.6 * \left(\frac{total\ syllables}{total\ words}\right) \quad (1)$$

The Flesch-Kincaid School Placement Test evaluates textbooks based on U.S. school placement. Because we are looking for an index that fits the Spanish language, it is rejected [3]. The formula for this test is mentioned in Eq. 2.

$$0.39 * \left(\frac{total\ words}{total\ sentences}\right) - 11.8 * \left(\frac{total\ syllables}{total\ words}\right) - 15.59 \quad (2)$$

The Flesch-Szigriszt readability index is an adaptation to the Spanish language of the Flesch index, mentioned above [3]. The formula for this test is mentioned in Eq. 3.

$$206.835 - 62.3 * \left(\frac{total\ words}{total\ sentences}\right) - \left(\frac{total\ syllables}{total\ words}\right) \quad (3)$$

All the indexes mentioned above consider the total syllables per total words of the text, but it was considered more important to study readability in terms of word length, since its simplification considers that short words imply a minor difficulty of understanding. The Gunning fog index uses this methodology and it is possible to adjust the formula. It is an index that indicates the readability of a text using a series of characteristics of it. In order to determine readability, the subject matter of the text is not considered. Common parameters are used, such as the number of syllables of the words that make

up the text. This index is established by an equation formulated by Robert Gunning in 1952 to identify the audience to which a given text can be directed. This method was created for the English language and is not as accurate for Spanish-language texts [4].

Although in Spanish the words are generally longer than in English, the frequency with which they appear in the texts in both languages is very similar, as seen in Fig. 1. This allows the Gunning Fog Index to be used to evaluate the readability of texts written in Spanish. However, since in Spanish longer words are used, a slight change was made in the Gunning formula. Through the Google Linguistic Corpus long words are chosen based on the frequency of occurrence in this and not based on its longitude. Obtaining as a result a more robust algorithm to measure readability in the Spanish language. The development of the methodology will be explained in later sections.

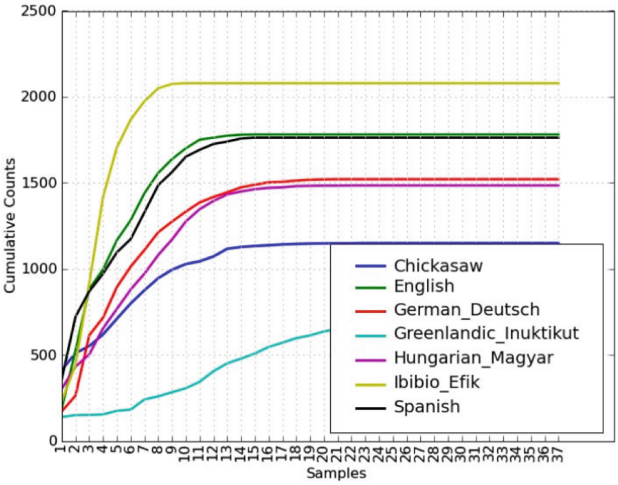

Fig. 1. Distribution of word length frequencies in different languages. Source: [4].

A distinction must be made between linguistic readability, which deals with verbal aspects, and typographical legibility, which refers to the visual perception of the text (layout of the text on the page, size of the letter, use of italics, bold, etc.). In this case, the linguistic readability will be evaluated, that is why some concepts must be known firsts.

The syllable is each of the phonological divisions into which a word is divided. It is the second smallest division of the spoken string. The phonological units into which a word is divided are called syllables, according to the minimum grouping of its articulated sounds, which means the union of a vowel and one or more consonants. In other words, these are the sound fragments into which a word can be divided, respecting the logic of its pronunciation [5].

The hiatus occurs when the accentuation of a word or its pronunciation forces to separate in different syllables a diphthong or a triphthong. This happens when there are two strong vowels, as well as when in a diphthong or in a triphthong the tonic vowel is a weak one [5].

A diphthong is a sound chain that is based on the articulation of two vowels, one followed by the other, without interruption and producing a smooth transition in the sound frequencies that characterize the timbres of each of the two vowels. Phonologically, two vowels articulated in this way are part of the same syllable. In a diphthong the acoustic formants have a smooth transition from one point of the vowel area to another, which gives them their diphthong nature. This is due to an articulation in which the tongue moves between different points during the emission of the diphthong. The two end points of the joint are perceived as the two vowels forming the diphthong. In the spectrogram of a hiatus the transition zone is not observed, that is why phonetically they are different [5].

The triphthong is the sequence of three vowels in the same syllable: closed vowel (u/i) + open vowel tonic (a/e/o) + closed vowel (u/i) [5].

Another fundamental concept is the Linguistic Corpus, because the implemented software makes use of it. A linguistic corpus is a broad and structured set of real examples of language use. These examples can be texts (the most common), or oral samples (generally transcribed) [6]. A linguistic corpus is a relatively large set of texts, created independently of their possible forms or uses. This means, in terms of its structure, variety and complexity, a corpus must reflect a language, or its mode, as accurately as possible; in terms of its use, concern that its representation is real. These corpora have similarities with the texts because they are composed of them, on the other hand, they are not texts in themselves, because unlike these, it does not make sense to analyze them in their entirety. A corpus lacks such characteristics because it does not have a structure, only a composition. For this reason, it is convenient to analyze a corpus using our own tools and methodology [6].

Each person writing a text forms his own style according to his knowledge and years of study. The selection of words and style will make a text clear, short and precise, or otherwise heavy or not very readable. On the other hand, a text must have unity, coherence and emphasis. Directly, the fog index does not measure the coherence of a text, because each person must evaluate if it is coherent under their own concept. The style is clear when it is readable. It is short when it does not contain useless words in the text. It is precise when a word cannot be removed without affecting the meaning of the sentence. Figure 2 shows how style relates to the quality of a text.

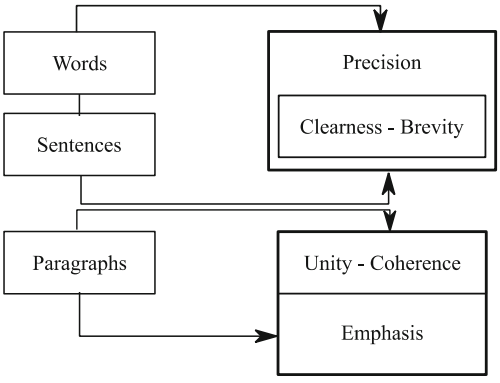

Fig. 2. Relation of writing style to words, sentences and paragraphs [7].

Readability tests were created to estimate the easy in which a text can be read, commonly expressed in years of study necessary to understand a text. Being able to measure the readability of a text allows educators to select texts that are relevant to their students by grade level, or their level of language proficiency for people who are learning a new language. In addition, readability tests help writers and publishers to verify that their texts meet the level of readability required for the target audience. Evaluating texts is increasingly important, due to the increasing variety, volume and complexity of written texts [8].

Since the popularization of personal computers, tools for writing analysis have been implemented. For example, the Word text editor performs document analysis if the user enables the option, in addition to correcting spelling, congruence and other errors. This tool uses the Flesch readability test and the Flesch-Kincaid grade level test. However, these are not enabled for the Spanish language [9].

Analysis of the readability of the texts has been used to verify that health-related materials have an appropriate level for most patients. For example, in [10] and [11] readability formulas are used to evaluate texts addressed to patients, so that thresholds can be established to guarantee the greatest readability for patients. Different hospitals, such as in [12] or in articles such as [13], have used readability tests to verify that documents of great importance, such as consent forms, have adequate readability for patients. Likewise, in web pages or health information documents, readability analyses have been carried out, finding that they are not optimal since they exceed the readability average [14]. Readability tests have also been applied in subjects related to: Parkinson's disease information [15], web-based cancer information [16], online educational materials for otolaryngology patients [17], orthopedic-related health topics [18], and others.

Readability tests have been used in conjunction with other tools to identify useful product reviews, due to the extensive proliferation of user-generated content, there has been a need to select useful information automatically. A wider range of classification characteristics is achieved through readability testing [19]. Because these same reviews are one of the main guides used by many users to decide which product to buy, they are often subject to manipulation, which is why in [20] used rating, readability and sentiment analysis to detect online review manipulation.

Related Works. The work of [22] describes a readability assessment approach to support the process of simplifying text for low-literate readers. Given an input text, the aim is to predict its level of readability, which corresponds to the level of literacy expected of the target reader, classified as rudimentary, basic or advanced. In this paper was explored the traditionally used characteristics plus the assessment of readability with several new characteristics and experimented with alternative ways of modeling the problem.

This article from [23] presents a different approach to the assessment of readability through classification. Through automatic learning a comparator is generated that judges the relative readability between two texts, and through this a set of given texts is ordered. The proposal solves the problem of the lack of training data, because the construction of the comparator only requires training data annotated with two levels of reading. The proposed method is compared with regression methods and a last generation classification method. An application called Terrace was developed, which retrieves texts with readability like that of a given input text.

In the work of [3], a tool was developed to analyze the readability of educational contents. The Gunning Fog Index was used to evaluate the texts. It was recommended to improve the formula that calculates the readability index in order to better adapt it to the Spanish language. On the other hand, in the work of [9], a Corpus in the Spanish language was used, constructed from Google Ngram and the Harvard university, the fog index was found by changing the long words for unknown words according to the frequency of appearance in the created corpus. So, the unknown words are calculated according to their frequency in the downloaded corpus.

2 Materials and Methods

2.1 Materials

In this research we used as materials a laptop with the following technical specifications: Intel core i5 processor, 8 Gb RAM and 1 Tb hard disk. The laptop must have internet access to be able to make requests to Google Ngram.

2.2 Methods

The fog index is an equation that measures the readability of an English text. It results in years of study necessary for a person to understand a text in a reading. Usually the fog index is used to adjust a text according to the level of the target audience. This idea was conceived by the American Robert Gunning, who considered that the use of long sentences and long words made it difficult to understand a text. Based on this, the fog index in paragraphs with a length of about one hundred words is calculated according to Eq. 4 [21]:

$$0.4\left[\left(\frac{N.\ words}{N.\ sentences}\right) + 100\left(\frac{N.\ long\ words}{N.\ word}\right)\right]\# \tag{4}$$

Where the terms:

- N. words represents the number of words in the paragraph.
- N. sentences is the number of sentences in the paragraph.
- N. long words corresponds to the number of long words in the paragraph.

A word is considered long, according to the Gunning Fog Index, if it has a length of three or more syllables. Except for proper pronouns, compound words and words that turn from three syllables when conjugated with English suffixes such as -ed, -es, or -ing [22].

Using Gunning fog index, you get a scale of values for English text, which correspond to: 5 as easy to understand, 10 more complicated, 15 difficult to understand and 20 very difficult to read. For example, most of the Bible has a fog index between 6 and 7. Magazines for the general public have an index of about 10. It should be noted that although a person with several years of study may understand a text with a fog index of 17 does not mean that it is pleasant to read these types of text.

Although the Gunning Fog Index is a practical and brief way to assess the complexity of compressing a text. It relates important characteristics of readability, such as sentence length and the use of long words. It has limitations in considering all long words as difficult. Since not all long words are complicated to understand. For example, the following long words are common and easy to understand in English, elephant, population, billion, etc. From the above, it can be deduced that the more common a long word is, less trouble it causes to the average reader [21].

The Zipf-Mandelbrot law was formulated in 1940 by George Kingsley Zipf. Establishing an empirical relationship between the frequency of a word, being inversely proportional to the nth word elevated to a value slightly greater than one [23]. So, the frequency of the second most repeated word will be about half of the first, the third word a third of the first, and so on. It is convenient to evaluate the frequency of the words in a logarithmic scale due to the non-linear variability between the frequency of a set of words.

Google Ngram Viewer, is a web page that shows a graph of the frequency of words separated by commas, using the annual count of N-grams in the different printed resources between 1500 and 2018, the Corpora has the languages: English, Spanish, Chinese, French, German, Hebrew, Italian and Russian. The algorithm can search by word or phrase, even if they have spelling errors or are meaningless [24].

The corpus of Google Books has limitations due to cultural popularity. One of its main problems is that the corpus is a library. It contains one of each book. So, a recognized author can significantly insert new words or phrases into Google Books vocabulary. Another problem lies in the inclusion of scientific texts, which have become an increasingly significant part of the corpus throughout the twentieth century. This results in phrases typical of academic articles, but less common at the general level [25].

3 Results

Initially, a program was implemented in Java to calculate the fog index according to Eq. 4. The algorithm enters a text into it and separates it into an array of words. The number of words is obtained according to the number of spaces and the number of sentences according to the number of points. Then, going through the arrangement of words is the number of long words, depending on whether the number of syllables of the word is greater than or equal to three. Finally, it is applied from the Gunning fog index parameters obtaining the readability of the text.

To separate the syllables a part of the algorithm in JavaScript was translated from [26] to Java, this code separates a word in syllables. Identifying correctly the formation of hiatuses, diphthongs and triphthong.

This program was applied for different texts in Spanish, of representative authors at world and local level, obtaining values greater than 17 for all. This shows that the index is not scaled for the Spanish language. According to the direct calculation of the fog index, much academic experience would be required to read texts such as "*100 years de soledad*" and "*El ingenioso hidalgo don Quijote de la Mancha*", which are texts commonly read in the literature area of colleges and universities. The comparative texts and their results can be seen in Table 1.

Table 1. Gunning fog index for different texts in Spanish.

Text	Author	paragraph	Fog index	Long words
El ingenioso hidalgo don Quijote de la Mancha	Miguel de Cervantes Saavedra	Cha 1/paragraph 1	22.35	54
100 años de soledad	Gabriel García Márquez	Cha 2/paragraph	24.75	67
El canto de las sirenas	William Ospina Buitrago	paragraph 2	27.88	96
Hace tiempo. Un viaje paleontológico ilustrado por Colombia	Carlos Jaramillo Muñoz y otros	Several	22.22	40
La Biblia		psalm 30	17.08	50

A clear limitation is that the Gunning fog index is not directly applicable to all languages. In the case of Spanish, words tend to be longer, resulting in much higher values for the fog index equation compared to the general fog index indicators. The book "*100 años de soledad*" would require approximately 25 years of study to understand the text in a single reading, which is equivalent to approximately one person with a post-doctorate.

A change was made in the Gunning equation, varying the long words to complex words. Considering the complex words which have a lower percentage of frequency of appearance compared to a defined threshold. To determine this threshold, a list of 60 words was made, classifying the words according to whether they were easy or complex. With the list of complex words, it was determined that the frequency value corresponded to 0.00015%, establishing this value for the threshold. Figure 3 below shows the words classified on a logarithmic scale, based on the threshold.

In Java, an API was made to obtain data from the Google Ngram page. Through requests using the package "org.json", which allows light and language-independent data exchange, including the ability to convert between JSON and XML, HTTP headers, cookies and CDL. So, you can send data to a page and receive its response. To make requests to Google Ngram, it relied on the API developed by [14], in which an API was developed to make requests by making modifications to the URL of Google Ngram.

Through the API, a Software was implemented that processes the entered text. Obtaining the number of sentences, number of words, and the number of complex words. For this, a paragraph is divided into words and organized into an arrangement, filtering the language signs, such as punctuation marks, question marks, exclamation marks, quotation marks, etc. With this arrangement you get the number of words and by passing it through a filter, you remove the words that are in the word buffer, which is constantly updated with the words that have been searched before. This buffer was used because the maximum number of words that can be done per query is 12 and generating many

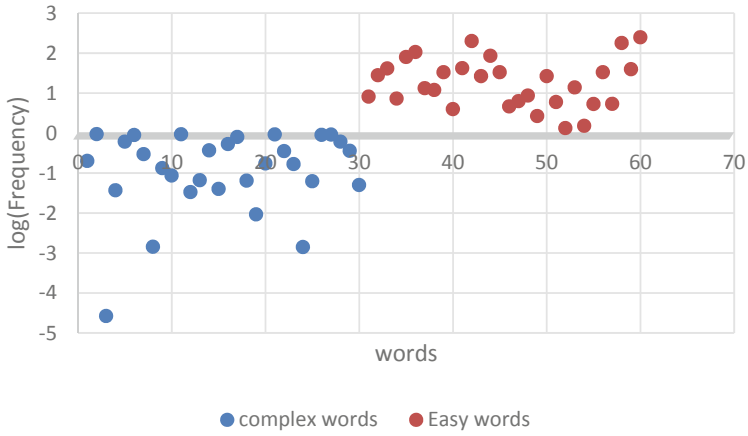

Fig. 3. Determination of the threshold for classifying words as complex or easy.

requests can make the page block the service momentarily. The algorithm scheme is shown in Fig. 4.

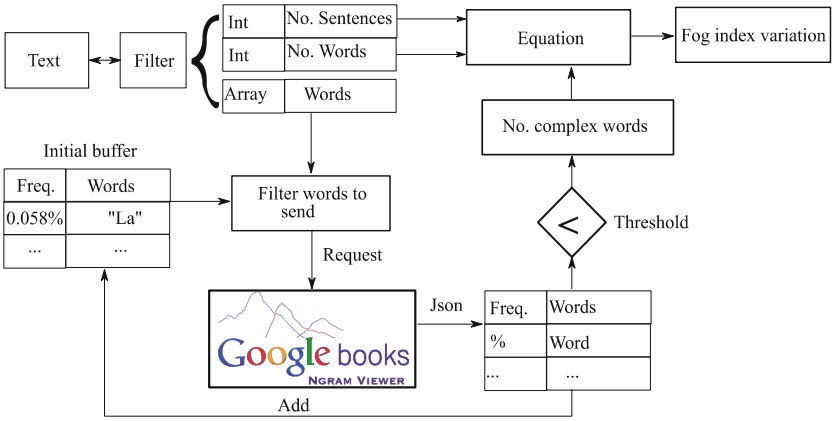

Fig. 4. Algorithm for calculating the fog index based on frequency.

The calculation of the fog index based on frequency was made for the same texts analyzed (with the Gunning fog index). Results were lower than the obtained using the Gunning fog index. On average the results differ by approximately 8 years, being a significant result, because they represent years of study equivalent to more than one university degree. Although the index decreases notably its value for the different texts, it is still a little high for the analyzed texts. The results can be seen in Table 2.

The slight increase in the fog index based on frequency is because Google Ngram tends towards academic literature. In order to verify the variation in the fog index, academic texts were collected from 10 university students over the course of their careers. They were asked to write at least two texts, the first in the first semesters of their careers,

Table 2. Comparison of the Gunning fog index and the variation performed.

Literary work	Fog index	Fog index based on frequency
El ingenioso hidalgo don Quijote de la Mancha	22.35	17.00
100 años de soledad	24.75	15.43
El canto de las sirenas	27.88	17.17
Hace tiempo. Un viaje paleontológico ilustrado por Colombia	22.22	11.76
La Biblia	17.08	12.44

and the second at the end of their careers. The criteria for the inclusion of the collected texts was that they had paragraphs of around 100 words. Texts of any academic subject, as exact sciences, humanities and thesis, were valid. In addition, the year or semester of completion of the text was requested. In total, all the students analyzed belonged to engineering careers. On average, the time difference in the elaboration of the evaluated texts was 6 semesters. The results obtained are shown in Table 3.

Table 3. Average score for the text at the beginning and end of the university course.

Text	Fog index	Long words	Fog index based on frequency	Complex words
1	29.25	37.90	18.62	5.10
2	27.55	51.40	15.08	7.10

Table 3 show that there was a decrease in the fog index and an increase in long words of text 1 to 2. However, it is confirmed that the directly applied Gunning fog index is not correct, because it does not have the appropriate scale. On the other hand, the fog index based on frequency showed an increase in readability in the students writing and an increase in the number of complex words used for writing, due to the process of university formation of each of these. This decrease can be explained due to the improvement of the style of each student, making more precise, coherent texts and making correct use of punctuation marks.

4 Discussions

Due to the boom in scientific literature on Google Ngram some words that are not considered difficult decrease in frequency. Some words in Spanish had a much lower frequency than the threshold, e.g.: "*escondiste*", "*clamé*", "*cantad*", etc. Although these words are not quotidian are easy to understand, but because Google Ngram has a more academic tendency they decrease in frequency. Academic texts often speak in the third person

and use words that are more focused on scientific contexts. Therefore, the developed algorithm was only tested in academic contexts.

As a future work it is expected to make a comparison of all the existing readability tests adapted to the Spanish language. In addition, it is planned to develop a tool to evaluate the readability of a text for the writing of essays and scientific articles. As well as allowing to measure the level of writing of a student, to verify that he has the necessary tools to start his university education process, and then follow its evolution.

5 Conclusions

A software was implemented to determine the readability of academic documents written in Spanish. Through a variation of the Gunning fog index using the Google Ngram linguistic corpus. It was possible to measure the readability level of different literary texts in Spanish. The results were verified with literary texts and documents made by university students at the initiations and finals of their careers. An average improvement of 3.54 in textual readability and an average increase of 2 complex words was evidenced.

Having a tool that allows quantifying the readability of documents written in the Spanish language, allows to generate processes to improve the written production of academic texts, helping to improve the readability of documents and their reception, by verifying that they meet the required level according to the target audience. On the other hand, knowing the level of readability of a text allows teachers to select the most appropriate documents for their students, according to their years of study or level of language proficiency, in the case of foreign students who are learning Spanish.

It is important to offer tools to improve the writing conditions of students, also improve the readability of academic institutions such as universities and publishing houses. Institutions that have as their main measurement standards the publication of scientific articles and literary texts, and deal with different levels of education.

References

1. Ferrando Belart, V.: La legibilidad: un factor fundamental para comprender un texto. Aten. Primaria **34**(3), 143–146 (2004)
2. Brucker, C.: Arkansas tech writing. English **2053**(June), 109 (2009)
3. Mata San Juan, H.: Herramienta de análisis de legibilidad de contenidos educativos. Graduate theses, Departamento de Informática, Universidad Carlos III de Madrid (2017)
4. Frías Delgado, A.: Distribución De Frecuencias De La Longitud De Las Palabras En Español. A Survey of corpus-based Research, pp. 756–770 (2009)
5. Coserio, E.: Introducción a la Lingüística (1983). https://www.csub.edu/~tfernandez_ulloa/spanishlinguistics/introduccion%20a%20la%20linguistica%20general.pdf
6. Pitkowski, E.F., Vásquez Gamarra, J.: Tinkuy Boletin de investigación y debate. El uso los corpus lingüísticos como Herram. pedagógica para la enseñanza y Aprendiz. ELE*, no. 11, pp. 31–51 (2005)
7. Mac Lean, A.: Comunicacion Escrita, pp. 5–17 (1975)
8. Zamanian, M., Heydari, P.: Readability of texts: state of the art. Theory Pract. Lang. Stud. **2**(1), 43–53 (2012)

9. Felipe Ovares, B., José Alberto, R.B.: Variación del Índice de Niebla Usando un Corpus Obtenido a Partir de los Libros Digitalizados por Google (2010)
10. Barrio-Cantalejo, I.M., Simón-Lorda, P., Melguizo, M., Escalona, I., Marijuán, M.I., Hernando, P.: Validación de la escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes. An. Sist. Sanit. Navar. **31**(2), 135–152 (2008)
11. Wang, L.W., Miller, M.J., Schmitt, M.R., Wen, F.K.: Assessing readability formula differences with written health information materials: application, results, and recommendations. Res. Soc. Adm. Pharm. **9**(5), 503–516 (2013)
12. Navarro-Royo, C., Monteagudo-Piqueras, O., Rodríquez-Suárez, L., Valentín-López, B., García-Caballero, J.: Legibilidad de los documentos de consentimiento informado del hospital La Paz. Rev. Calid. Asist. **17**(6), 331–336 (2002)
13. Simón Lorda, P., Barrio Cantalejo, I.M., Carro, L.C.: Legibilidad de los formularios escritos de consentimiento informado (1996)
14. Blanco Pérez, A., Gutiérrez Couto, U.: Legibilidad de las páginas web sobre salud dirigidas a pacientes y lectores de la población general. Rev. Esp. Salud Publica **76**(4), 321–331 (2002)
15. Fitzsimmons, P.R., Michael, B.D., Hulley, J.L., Scott, G.O.: A readability assessment of online Parkinson's disease information. J. R. Coll. Phys. Edinb. **40**(4), 292–296 (2010)
16. Friedman, D.B., Hoffman-Goetz, L.: A systematic review of readability and comprehension instruments used for print and web-based cancer information. Heal. Educ. Behav. **33**(3), 352–373 (2006)
17. Svider, P.F., et al.: Readability assessment of online patient education materials from academic otolaryngology-head and neck surgery departments. Am. J. Otolaryngol. - Head Neck Med. Surg. **34**(1), 31–35 (2013)
18. Badarudeen, S., Sabharwal, S.: Assessing readability of patient education materials: current role in orthopaedics. Clin. Orthop. Relat. Res. **468**(10), 2572–2580 (2010)
19. O'Mahony, M.P., Smyth, B.: Using readability tests to predict helpful product reviews (2010)
20. Hu, N., Bose, I., Koh, N.S., Liu, L.: Manipulation of online reviews: an analysis of ratings, readability, and sentiments. Decis. Support Syst. **52**(3), 674–684 (2012)
21. Seely, J.: Oxford Guide to Effective Writing and Speaking: How to Communicate Clearly (2013)
22. Aluisio, S., Specia, L., Gasperin, C., Scarton, C.: Readability Assessment for Text Simplification (2010)
23. Montemurro, M.A.: Beyond the Zipf-Mandelbrot law in quantitative linguistics. Phys. A **300**, 567–578 (2001)
24. Michel, J.-B., et al.: Quantitative analysis of culture using millions of digitized books. Science **331**(6014), 176–182 (2011)
25. Pechenick, E.A., Danforth, C.M., Dodds, P.S.: Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. PLoS ONE **10**(10), e0137041 (2015)
26. Cofré, N., Arce, J.: Librería para obtener las silabas, posición de la silaba tónica, tipo acentuación, hiato, diptongo y triptongo de una palabra. https://github.com/ncofrem/silabajs. Accessed 14 Oct 2019
27. Fisher, J.: API for Google Ngram Viewer. https://jameshfisher.com/2018/11/25/google-ngram-api/. Accessed 14 Oct 2019