Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 127-130 recibido 28-03-2020 revisado 30-05-2020 aceptado 30-05-2020

The Text Complexity Library

Biblioteca de Complejidad Textual

Rocío López-Anguita, Jaime Collado-Montañez, Arturo Montejo-Ráez
Centre for Advanced Information and Communication Technologies (CEATIC)
University of Jaén, Spain
{rlanguit, jcollado, amontejo}@ujaen.es

Abstract: This paper introduces a new resource for computing textual complexity. It consists in a Python library for calculating different complexity metrics for several languages from plain texts. The resource has been made available to the research community and provides all needed instructions for its installation and use. To our knowledge, it is the first time a resource like this is published, so we expect many researchers can profit from it.
Keywords: Demostration, linguistic resources, textual complexity, lexical analysis

Resumen: Este artículo presenta un nuevo recurso para el cálculo de la complejidad textual. Se trata de una biblioteca de programación en Python que facilita el cómputo de distintas métricas de complejidad para varios idiomas a partir de textos en lenguaje natural. El recurso se ha liberado para su uso por parte de la comunidad científica y proporciona todas las instrucciones necesarias para su instalación y aprovechamiento. Hasta donde sabemos, es la primera vez que un recurso así está disponible, por lo que esperamos sea de utilidad.
Palabras clave: Demostración, recursos lingüísticos, complejidad textual, análisis léxico

1 Introduction

Reading comprehension and reading competence are complex processes that are closely related, according to Pérez (2014), as is the concept of readability. The reading comprehension is close to the reader's capabilities and the latter is an objective view of the complexity of the text.

Determining the readability of a text is not a simple task, as each reader has different skills or limitations (Cain, Oakhill, and Bryant, 2004). It is usually determined by linguistic features, which are usually grouped into those related to grammar (in other words, syntax) and those related to the lexicon (i.e. vocabulary) (Alliende González, 1994).

Currently, we consider that measures of complexity can be a convenient way to model natural language in certain applications, such as authorship detection, text selection for people with difficulties associated with language disorders (autism, cerebral palsy...), or early detection of cognitive impairments, such as Alzheimer's. Therefore, in this article we present a "demo" paper, which consists of 12 of the most widely used metrics for lexical and syntactic readability. These measures and their interpretation are presented below, as well as details on the use of the library.

2 Complexity metrics provided

In this section, we introduce the different complexity metrics offered in this Python library, proposed by different authors, for different languages (Spanish, English, French...).

Lexical complexity: The lexical complexity of a text, determined by the frequency of use and lexical density, was proposed by Anula (2008). It is based on the number of different content words per sentence (*Lexical Complexity Index, LC*) and on measuring the number of low frequency words per 100 content words (*Index of Low Frequency Words, ILFW*). Consequently, the higher the LC index, the greater the difficulty in reading comprehension.

Spaulding readability: Commonly known as the SSR Index, it was proposed by Spaulding (1956). It focuses on measuring vocabulary and sentence structure to predict

ISSN 1135-5948. DOI 10.26342/2020-65-19

the relative difficulty of a text's readability. Its formula is an empirically adjusted measure to try to keep the score between 0 and 1.

Complexity of sentences: The Sentence Complexity Index (SCI) was proposed by Anula (2008), as a measure of the complexity of sentences in a literary text aimed at second language learners.

This syntactic complexity measure focuses on measuring the number of words per sentence, thus obtaining the sentence length index (*Average Sentence Length, ASL*), and the number of complex sentences per sentence, from a complex sentence index (*Complex Sentences, CS*).

Automated Readability Index (ARI): Senter and Smith (1967) proposes one of the most used indexes due to its ease of calculation, the Automated Readability Index, better known as ARI (*Automated Readability Index*). This index measures the difficulty of a text from the average number of characters (letters and numbers) per word and the average number of words per sentence.

Dependency tree depth: This measure was proposed by Saggion et al. (2015). It is a very useful metric to capture syntactic complexity: long sentences can be syntactically complex or contain a large number of modifiers (adjectives, adverbs or adverbial phrases). It complements the ASL measure, as it captures syntactic complexity in terms of recursive or nested structures.

Punctuation Marks: This measure was also proposed by Saggion et al. (2015). In the complexity of a text, the average number of punctuation marks is used as one of the indicators of the simplicity of the text.

Readability of Fernández-Huerta: Blanco Pérez and Gutiérrez Couto (2002) y Ramírez-Puerta et al. (2013) propose this measure of complexity as an adaptation to Spanish of Flesch's readability test (Flesch (1948)).

Readability of Flesch-Szigrist (IFSZ): The works of Barrio-Cantalejo et al. (2008) and Ramírez-Puerta et al. (2013) propose the Flesch-Szigristzt readability index as a modification of the Flesch formula (Flesch, 1948) adapted to Spanish by Szigriszt-Pazos in 1993. This index is currently considered a reference for the Spanish language. It focuses on measuring the number of syllables per word and the number of sentences per word in the text.

Comprehensibility of Gutiérrez de Polini: This metric, originally developed in 1972, is not an adaptation of English, but was created from the beginning for Spanish (Rodríguez, 1980). It focuses on measuring the average number of letters per word and the average number of words per sentence.

μ Readability: It is a formula to calculate the readability of a text. It provides an index between 0 and 100 and was developed by Muñoz (2006). This measure focuses on measuring the number of words, the average number of letters per word and their variance.

Minimum age to understand: In work of García López (2001) we can find another formula to measure the age needed to understand a text. It is, again, an adaptation into Spanish of Flesch's original formula (Flesch (1948)) for English. It measures the average number of syllables per word and the average number of words per sentence to obtain the minimum age needed to understand a text.

SOL Readability: Contreras et al. (1999) proposes the SOL metric as an adaptation to Spanish of the SMOG formula proposed by Mc Laughlin (1969). It measures the readability of a text by means of grade level, which is the number of years of schooling required to understand the text.

Years Crawford: This measure was proposed by *Alan N. Crawford* in 1989 (Crawford, 1984). It is used to calculate the years of school required to understand a text. Measures the number of sentences per hundred words and the number of syllables per hundred words.

3 How to obtain it

The library has been released under the General Public License (GPL v3.0) license[1] and can be downloaded or cloned from its public repository[2]. The library will be updated with new features in the future, and you can always get the latest version from that link.

[1] `http://www.gnu.org/licenses/gpl.html`

[2] `https://gitlab.ujaen.es/amontejo/text-complexity`

4 Installation

In order to use this library, you first need to install some previous requirements:

- NumPy, Scipy, Pandas, Matplotlib and Openpyxl for python3 have to be installed in your system.
- The FreeLing (Padró and Stanilovsky, 2012) package, which is a library providing language analysis services that our library makes use of. In order to install it, you have to follow its installation manual under the project's GitHub page[3]

5 Usage examples

In order to test the library and teach how to use it, we have prepared some testing texts for Spanish under the `./texts` folder (you can use your own texts by modifying these files or adding more to that location).

To compute complexity metrics on these text samples, modify the `FREELINGDIR` variable in the `TextComplexityFreeling.py` script (line 18) to your own FreeLing installation directory (`/usr/local` by default).

Then, if you run the Jupyter notebook `examples.ipynb` you should get some tables with the metrics for each text provided. The script will also generate three MS Excel files containing the results in your project's folder.

To use it in your Python scripts, this is as simple as follows:

```
import TextComplexityFreeling as TCF
# Create the text complexity calculator
tc = TCF.TextComplexityFreeling()
# Load text to analyze
text_processed = tc.textProcessing(text)
# Compute different metrics
pmarks = tc.punctuationMarks()
lexcomplexity = tc.lexicalComplexity()
ssreadability = tc.ssReadability()
sencomplexity = tc.sentenceComplexity()
autoreadability = tc.autoReadability()
embeddingdepth = tc.embeddingDepth()
readability = tc.readability()
agereadability = tc.ageReadability()
yearscrawford = tc.yearsCrawford()
```

For example, for a given sample text:

La última luna llena del año, que se observará completa este jueves en el cielo, será especial. Se produce estos días el fenómeno conocido como luna fría, una coincidencia astronómica que hace las delicias de quienes atribuyen al astro cualidades esotéricas. Sucede cuando la Tierra se encuentra ubicada exactamente entre el sol y la luna, de forma que la luna recibe directamente la luz. La luna llena será visible durante toda la noche, pero alcanzará su magnitud máxima cuando se encuentre a medio cielo, de forma que, al reflejar completamente la luz del sol que incide en la tierra, se verá especialmente grande y luminosa. Se llama luna fría porque marca la llegada del invierno en el hemisferio norte, aunque también se conoce como luna de las noches largas al ocurrir cerca del solsticio, informa National Geographic, que cita a un astrónomo de la NASA. La luna fría de 2019 coincide además con la lluvia de meteoros Gemínidas, visible entre el 7 y el 17 de diciembre pero que alcanzará su punto máximo de actividad entre el 11 y el 13. Es la lluvia de estrellas más masiva, lo que la hace mucho más brillante. El cielo augura todo un espectáculo esta noche.

This is the generated output:

```
Number of words (N_w):  207
Punctuation marks: 21
Number of low freq. words (N_lfw):  67
Number of content words (N_dcw):  71
Number of sentences (N_s):  8
Number of total content words (N_cw):  93
Lexical Distribution Index (LDI): 8.875
Index Low Frequency Words (ILFW):  0.72
Lexical Complexity Index (LC): 4.7977
Number of rare words (N_rw):  65
Spaulding Spanish Readability (SSR): 167.82
Average Sentence Length (ASL): 25.875
Complex Sentences (CS): 23.75
Sentence Complexity Index:(SCI): 24.81
Automated Readability Index (ARI): 13.66
Average embddings depth (MeanDEPTH): 7.25
Huerta's Readability index:  80.86
IFSZ Readability: 54.25
Polani's Compressibility (Polani's): 42.61
Mu Readability: 53.19
Minimum age to understand: 12.48
SOL Readability: 11.66
Years needed: 5.76
```

6 Conclusions and future versions

There are several studies that reflect the strong influence of the richness of the reader's vocabulary on reading comprehension, beyond symbols and grammar. However, we consider that complexity measures can be a convenient way to model natural language in

[3] `https://github.com/TALP-UPC/FreeLing-User-Manual`

certain applications, such as authorship detection, text selection for people with difficulties associated with language disorders (autism, cerebral palsy...), or early detection of cognitive impairments, such as Alzheimer.

Another future line of work is to define complexity from computed language models (like RNNs or BERT models). We believe that information measures on the parameters of these models may capture the inherent complexity of the texts they were trained on.

7 Acknowledgements

This work has been partially supported by Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government.

References

Alliende González, F. 1994. La legibilidad de los textos. *Santiago de Chile: Andrés Bello*, 24.

Anula, A. 2008. Lecturas adaptadas a la enseñanza del español como l2: variables lingüísticas para la determinación del nivel de legibilidad. *La evaluación en el aprendizaje y la enseñanza del español como LE L*, 2:162–170.

Barrio-Cantalejo, I. M., P. Simón-Lorda, M. Melguizo, I. Escalona, M. I. Marijuán, and P. Hernando. 2008. Validación de la Escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes. In *Anales del Sistema Sanitario de Navarra*, volume 31, pages 135–152. SciELO Espa{ na.

Blanco Pérez, A. and U. Gutiérrez Couto. 2002. Legibilidad de las páginas web sobre salud dirigidas a pacientes y lectores de la población general. *Revista española de salud pública*, 76(4):321–331.

Cain, K., J. Oakhill, and P. Bryant. 2004. Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of educational psychology*, 96(1):31.

Contreras, A., R. Garcia-Alonso, M. Echenique, and F. Daye-Contreras. 1999. The sol formulas for converting smog readability scores between health education materials written in spanish, english, and french. *Journal of health communication*, 4(1):21–29.

Crawford, A. N. 1984. A spanish language fry-type readability procedure: Elementary level. bilingual education paper series, vol. 7, no. 8.

Flesch, R. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

García López, J. 2001. Legibilidad de los folletos informativos. *Pharmaceutical Care España*, 3(1):49–56.

Mc Laughlin, G. H. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Muñoz, M. 2006. Legibilidad y variabilidad de los textos. *Boletín de Investigación Educacional, Pontificia Universidad Católica de Chile, 21*, 2:13–26.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2473–2479.

Pérez, E. J. 2014. Comprensión lectora vs competencia lectora: qué son y qué relación existe entre ellas. *Investigaciones sobre lectura*, (1):65–74.

Ramírez-Puerta, M., R. Fernández-Fernández, J. Frías-Pareja, M. Yuste-Ossorio, S. Narbona-Galdó, and L. Peñas-Maldonado. 2013. Análisis de legibilidad de consentimientos informados en cuidados intensivos. *Medicina Intensiva*, 37(8):503–509.

Rodríguez, T. 1980. Determinación de la comprensibilidad de materiales de lectura por medio de variables lingüísticas. *Lectura y vida*, 1(1):29–32.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14.

Senter, R. and E. A. Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Spaulding, S. 1956. A spanish readability formula. *The Modern Language Journal*, 40(8):433–441.